VMA: Domain Variance- and Modality-Aware Model Transfer for Fine-Grained Occupant Activity Recognition

Zhizhang Hu
zhu42@ucmerced.edu
Univ. of California Merced
Merced, California, USA

Yue Zhang
yzhang58@ucmerced.edu
Univ. of California Merced
Merced, California, USA

Tong Yu
tyu@adobe.com
Adobe Research
San Jose, California, USA

Shijia Pan
span24@ucmerced.edu
Univ. of California Merced
Merced, California, USA

ABSTRACT

The growth of the Internet of Things (IoT) sensing systems leads to a large number of multimodal datasets over different deployments. Labeling costs for these datasets, especially fine-grained labels, are often tremendous. On the other hand, different data distributions (domain variance) of these datasets prevent models built with labels of one dataset (source domain) from being directly used in another (target domain). This domain variance may be caused by one or more physical factors change in the deployments, such as buildings and/or people. Existing model transfer studies mainly focus on adapting the model to the domain variance caused by only one physical factor change. When multiple factors change between the source and target domains, the model transfer often yields low accuracy due to significant domain variance.

We present *VMA*, a model transfer framework for multimodal IoT sensing data that handles multi-factor domain variance. *VMA* first decouples the multi-factor domain variance between two datasets to multiple single-factor domain variance dataset pairs with other available datasets. Then, *VMA* leverages sensing modalities robust to each single-factor domain variance for accurate prediction by weighing them more in the fusion. We apply *VMA* to the fine-grained occupant activity recognition application with a multimodal sensing system of structural vibration and wearable IMU. We collect real-world datasets to evaluate the proposed framework. *VMA* achieves a model transfer accuracy up to 76.1% on the target domain with multi-factor domain variance, demonstrating a 1.6× and 1.9× error reduction compared to direct prediction baselines with and without modality-aware learning design.

KEYWORDS

Multimodal Sensing, Multi-factor Domain Variance, Model Transfer

1 INTRODUCTION

Internet of Things (IoT) systems enable various smart building applications such as in-home older adults/patient monitoring via multimodal sensing [19, 22, 24]. However, one of the bottlenecks that limits the scalability of these IoT sensing systems is the cost of labeling, especially for fine-grained labels [13]. This is because datasets collected at different deployments often have different data distributions. We define the multimodal dataset follows one data distribution as one domain, and the difference between each dataset's distribution is therefore referred to as the domain variance.

To mitigate the intensive needs of labeled data, many model transfer studies have been conducted, such as adversarial deep domain adaptation [42] and domain-invariant feature learning [26]. Such approaches have been explored to reduce the labeled data needed when the physical factor like occupant [42], structure [20], device [1], illumination [39] changes and induce domain variance between training (source domain) and test datasets (target domain). These works focus on one physical factor change-induced domain variance. When there are multiple physical factor changes between two datasets, the domain variance is more significant than that of single factor changes, which leads to low prediction accuracy.

We propose *VMA*, a domain variance- and modality-aware model transfer framework to handle multi-factor domain variances efficiently. The intuitions are twofold. Firstly, for source and target domains with multi-factor domain variances, there are other domains that are of single-factor variance to them. *VMA* **decouples the multi-factor domain variance to a transfer path** of multiple single-factor domain variances. We refer to datasets on this path as intermediate domains. Secondly, for a single-factor difference between datasets, some sensing modalities would have a less significant domain variance than other modalities, meaning their model transfer often yield a higher accuracy. *VMA* conducts a **modality-aware model transfer** along the transfer path by first learning a multi-task model with modality-specific input and output layers to predict the intermediate domain data. Then *VMA* selects predictions with high confidence to pseudo-label the intermediate domain data. Then the pseudo-labeled intermediate domain data is used to train another multi-task model for model transfer. The model transfer is done by reusing shared hidden layers' parameters to retain the modality-invariant knowledge. We apply *VMA* on the application of fine-grained occupant activity recognition [13] as a demonstration. We select structural vibration-based human sensing and wearable IMU as the representative sensor modalities. They provide complementary spatiotemporal information of the occupant activities. The contributions of this work are as follows.

- We present *VMA*, a model transfer framework for multimodal IoT sensing data that handles multi-factor domain variance.
- We introduce a domain variance decoupling algorithm that generates a transfer path leveraging physical knowledge.
- We demonstrate *VMA* with the fine-grained occupant activity recognition application with a multimodal sensing system of structural vibration and wearable IMU.
- We evaluate *VMA* with data collected from real-world and compare system performance over various baselines.

The rest of this paper is organized as follows. Section 2 analyzes modalities' sensitivity to domain variance of different factors. Then, Section 3 presents the system design and the domain variance- and modality-aware model transfer algorithm. Next, Section 4 shows the experiments and evaluation analysis. Section 5 lists prior work

978-1-6654-9624-7/22/$31.00 ©2022 IEEE
DOI 10.1109/IPSN54338.2022.00028

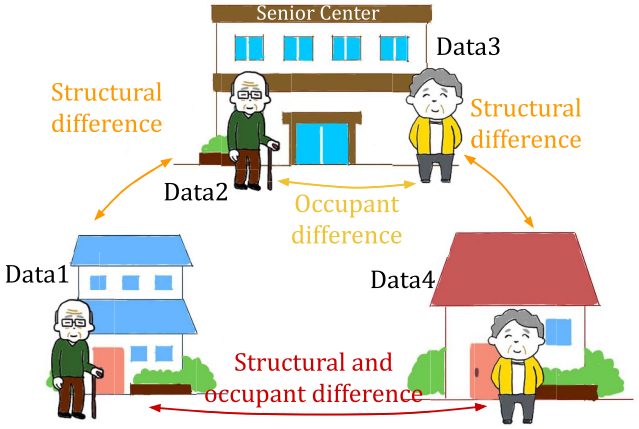

Figure 1: Examples of multi-factor and single-factor domain variances. Datasets are collected for two occupants in different structures. For the same occupant at different structures, i.e., Data1/Data2 and Data3/Data4, the datasets are of single-factor domain variance caused by the structure difference. For different occupants at the same structure, i.e., Data2/Data3, the datasets are of single-factor domain variance caused by the occupant difference. For Data1/Data4, both structure and occupants are different, therefore they are of multi-factor domain variance.

related to our system and algorithm. Finally, Section 6 discusses the potential future directions and Section 7 concludes this work.

2 MODALITY SENSITIVITY ANALYSIS

Different sensors acquire occupant information via different sensing principles. We analyze the sensing principles for structural vibration sensors and wearable IMU, to depict the key factors that directly impact the acquired signal/data. The analysis is used as the metric to determine the order of model transfer in Section 3.3.

2.1 Structural Vibration Sensing

Structural vibration sensors capture vibrations induced by occupant activities. When an occupant interacts with a surface, the interaction induces a surface deformation. This deformation generates vibration. This vibration propagates as a wave in the structure from interact location to sensor location. Finally, the sensor captures this vibration and converts the motion of the surface to voltage. Therefore, the key factors that impact the data property are threefold.

1) vibration generation. We use a single degree-of-freedom system with properties represented by mass, spring, and damper to simplify the surface vibration [40]. The external force F applied to the system at time t can be described as:

$$F(t) = ma(t) + cv(t) + kz(t) = m\ddot{z}(t) + c\dot{z}(t) + kz(t) \quad (1)$$

Where m, k, and c are the mass, spring constant, and damping coefficient. a, v, and z are acceleration, velocity, and displacement respectively. By solving Eq. 1, we can acquire the displacement of the induced vibration wave $z_s(t)$. For the same external force $F(t)$, different parameter (m, k, and c) values of different structures would result in different vibration wave $z_s(t)$. For example, concrete-steel floor often have higher stiffness (larger k) [27], larger density (larger m) [2] compared to wooden floors.

2) wave propagation. It is formulated as a wave attenuation model of the path between excitation and the sensor [7]:

$$z_r(d) = \frac{z_s}{d\sqrt{d}} e^{\alpha d} \quad (2)$$

where $z_r(d)$ is the received vibration at distance d, z_s is the source vibration, and α is the material-dependent attenuation coefficient. The attenuation coefficient is determined by the structural properties [33] and impacts waveforms when vibrations propagate.

3) signal acquisition. Different types of sensors measure vibrations in different forms. For example, the geophone sensor measures the velocity $\dot{z_r}$. The accelerometer measures the acceleration $\ddot{z_r}$. Therefore, the same vibration z_s will have different waveforms and spectrum characteristics when captured by different sensors. In this paper, we do not consider this impact because it can be calibrated by manufacturer. In summary, the vibration signal z_r is directly impacted by the structure parameters based on Eq. 1 and Eq. 2.

2.2 Wearable IMU Sensing

Wearable IMUs are attached to and measure body parts' motions.

1) signal generation. We use a 3D rotation model to simplify the body part movement motion [34], assuming that the joint is fixed and the target body part rotates. For example, when a person moves their hand, we consider the shoulder is a fixed point and the arm length is the rotation radius. For a given motion with displacement $s(t)$, the linear velocity $v(t)$ and angular velocity $\omega(t)$ of the moving body part at time t can be represented as:

$$v(t) = \dot{s}(t), \quad \omega(t) = v(t)/l = \dot{s}(t)/l \quad (3)$$

where l is the distance between the body part and the joint.

2) signal acquisition. The IMU sensor directly measures motions in the form of linear acceleration and angular velocity [14]. For a motion with a displacement of $s(t)$, the IMU outputs linear acceleration $\ddot{s}(t)$ and angular velocity $\omega(t) = \dot{s}(t)/l$. As Eq. 3 shows, the change of l and $\dot{s}$ would result in different linear acceleration and angular velocity. For example, older adults usually have a lower $\dot{s}$ and children often have a shorter l. Therefore, the IMU data is directly impacted by occupants' physical and motion parameters.

In summary, from the data acquisition perspective, the vibration signal/data is sensitive to structural differences, and the IMU signal/data is sensitive to occupant variation.

3 SYSTEM DESIGN

VMA aims to tackle the challenge of multi-factor domain variance in real-world multimodal IoT sensing datasets. Figure 2 depicts *VMA* and its four modules with the fine-grained activities recognition application. We adopt structural vibration and on-wrist IMU sensors to capture spatiotemporal characteristics of fine-grained occupant activities (Section 3.1). This multimodal data is then sent to the modality-aware data characterization module (Section 3.2), where *VMA* 1) determines the spatial characteristics of the given data and 2) extracts signal features. Simultaneously, *VMA* finds the transfer path between the source and target domain in the domain variance-aware multi-factor decoupling module (Section 3.3). Finally, *VMA* conducts modality-aware model transfer with inputs of multimodal signal features of datasets on the transfer path (Section 3.4).

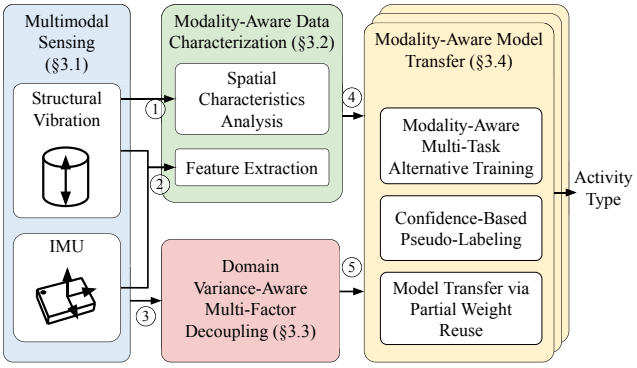

Figure 2: *VMA* framework with four modules for the application of fine-grained activity recognition 1) multimodal sensing, 2) modality-aware data characterization, 3) domain variance-aware multi-factor decoupling, and 4) modality-aware model transfer. The data flows marked in the figure are respectively: ① structural vibration signal segments of detected events; ② multimodal signals; ③ sensing system meta data, including structure ID, occupant ID (wearable ID), labeled/unlabeled; ④ multimodal signal features grouped by the same spatial characteristics (area ID). ⑤ decoupled transfer path between source and target domains.

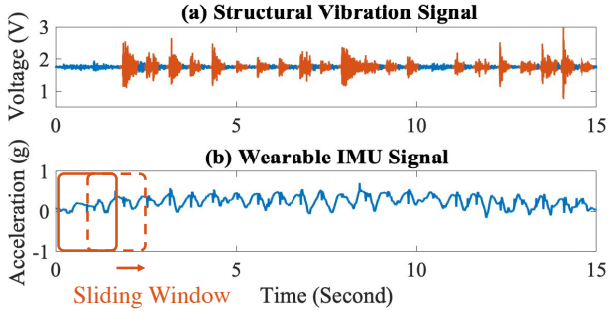

Figure 3: Exemplary signals of structural vibration and wearable IMU based occupant sensing. (a) vibration signal with detected events marked in red lines. (b) one axis of the acceleration signals with sliding window depicted in red boxes.

3.1 Multimodal Sensing System

Structural vibration and IMU sensors demonstrate complementarity in capturing spatiotemporal information [13]. Similar to [13], we utilize structural vibration and IMU sensors to predict fine-grained occupant activities in our system.

3.1.1 Structural Vibration Sensing. Structural vibration sensors capture occupant activities when they interact with ambient surfaces and induce the vibration (Section 2.1). We place vibration sensors on surfaces (e.g., table, countertop, floor) over different areas (e.g., kitchen, study) to capture vibrations induced by different types of activities. These activities often induce impulsive vibration signals, which we define as **event**s, shown as red solid lines in Figure 3 (a). We apply a sliding window to the acquired sensor signal. We establish a Gaussian noise model with the energy of the windowed signal when no events occur. Then we conduct anomaly detection on the incoming windowed signal based on this Gaussian noise model [20]. The windows with signal energy detected as anomalies are considered as events. If consecutive windows are detected as anomalies, they are considered to be the same event, i.e., the length of the event may vary over different activity types. For a structure deployed with N vibration sensors, if one sensor detects an event, *VMA* considers it an event for all N sensors. Further processing and learning are done on the **event level**.

3.1.2 Wearable Sensing. We use an on-wrist IMU sensor to capture the motion of participants' dominant hand to infer the type of activities. Since the IMU sensor captures the motion of the attached body part all the time, the concept of 'event' defined for vibration data is not suitable for IMU data. We adopt the sliding window to segment the signal into windows of size w_{IMU}, as depicted in red boxes in Figure 3 (b), and predict the current activity at the **window level**.

3.2 Multimodal Data Characterization

Given a dataset from an unknown deployment, *VMA* pre-processes the dataset by extracting the following information.

3.2.1 Spatial Characteristics Analysis. To augment the spatial information for activity recognition, we conduct spatial characteristics analysis on the infrastructural sensing – structural vibration sensors. Vibration sensors detect activities within the sensing range and are deployed over multiple areas (e.g., kitchen and study). For a deployment with N sensors covering n areas, *VMA* trains a multimodal multi-task learning model for each area, i.e., n models in total. Meanwhile, *VMA* keeps track of a system status flag F_{Area} indicating the event area. When events are detected by vibration sensors, the sensor with the highest event signal energy is considered as the closest to the activity, and we set the F_{Area} as the area ID. The data points with the same F_{Area} values are trained/tested with the area-specific model.

3.2.2 Feature Extraction. For vibration signals, *VMA* extracts frequency components as features of the detected events. For the IMU sensor, we apply a sliding window on six axes (accelerometer and gyroscope). For each window, *VMA* extracts 36 key features from each axis [13, 35] and then concatenates features from all axes.

3.3 Domain Variance-Aware Multi-Factor Decoupling

In real-world datasets, the data distribution change between the labeled and unlabeled datasets is often caused by multiple factors in a coupled manner [25]. We assume the investigated datasets are impacted by r known domain variance factors (e.g., occupant, structure, device, illumination) and have data from q modalities. We denote a domain as $\mathbf{D} = [f_1, f_2, ..., f_r]$. For each pair of datasets, we quantify their domain variance using *Factor Difference* as $\mathbf{FD} = [FD_1, ..., FD_r] \in \mathbb{Z}_2^r, \mathbb{Z}_2 = \{0, 1\}$ to encode the domain variance factor difference between two datasets. $FD_i = 1$ means the two domains are different in the i^{th} factor. We also construct the *Shared Modality* as $\mathbf{SM} = [SM_1, ..., SM_q] \in \mathbb{Z}_2^q$, where $\mathbb{Z}_2 = \{0, 1\}$, to encode the shared modalities between two domains. $SM_i = 1$ means

the corresponding modality is available in both datasets. We then leverage the analysis in Section 2 and establish an *Factor-Modality Sensitivity Matrix* $\mathbf{FM} \in \mathbb{Z}_2^{q \times r}$, where $\mathbb{Z}_2 = \{0, 1\}$. In **FM**, each row stands for sensing modalities in the same order as the **SM**. Each column stands for domain variance factors in the same order as the **FD**. We assign $\mathbf{FM}_{i,j}$ as 0 when the i^{th} modality is directly impacted by the j^{th} factor. Otherwise, we assign $\mathbf{FM}_{i,j}$ as 1, indicating an indirect or less sensitive impact.

We consider datasets with **FD** that has more than one element of 1 are non-directly transferable, because they have multi-factor domain variance. For datasets with **FD** that has only one element of 1, i.e., single-factor domain variance, *VMA* calculates $\mathbf{Tr} = (\mathbf{FM} \cdot \mathbf{FD})^T \cdot \mathbf{SM}$, which counts the number of available modalities that are not directly impacted by this single-factor domain variance. If $\mathbf{Tr} > 0$, we consider the pair of datasets are **directly transferable**. For the example shown in Figure 1, there are two domain variance factors ($q = 2$) between Data 1 and Data 4, i.e., occupant and structure. Considering *VMA* with two modalities ($r = 2$), i.e., structural vibration and wearable IMU , we form $\mathbf{FM} = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$. The rows are IMU and vibration sensing. The columns are occupant and structure factors. Between Data 1 and Data 4, we form $\mathbf{FD} = [1, 1]$ and $\mathbf{SM} = [1, 1]$. Since there are more than one 1 in **FD**, their domains are not directly transferable. For Data 1 and Data 2, we form $\mathbf{FD} = [0, 1]$ and $\mathbf{SM} = [1, 1]$. We calculate $\mathbf{Tr} = (\mathbf{FM} \cdot \mathbf{FD})^T \cdot \mathbf{SM} = 1 > 0$, meaning their domains are directly transferable.

After calculating pair-wise **Tr** between all available domain pairs, *VMA* establishes a graph with domains as nodes and $\mathbf{Tr} > 0$ as edges. Then, *VMA* conducts a breadth-first search (BFS) to select intermediate domains between source and target domains. The labeled source domain is the starting point of the BFS. The graph connectivity is defined by 1) the pair-wise transferability, and 2) all selected domains should share at least two sensing modalities, to enable further model transfer. If BFS returns a path, *VMA* considers the model transfer feasible.We refer to this path as the **transfer path**, denoted as $\mathbf{D}_1 \rightarrow \mathbf{D}_2 \rightarrow ... \rightarrow \mathbf{D}_p$, where p is the path length and $\mathbf{D}_p$ is the target domain. If there are multiple paths returned by BFS, *VMA* adopts the first searched path. If no path is returned, *VMA* waits for more datasets.

3.4 Modality-Aware Model Transfer

Given a labeled or pseudo-labeled domain $\mathbf{D}_i$, *VMA* conducts the modality-aware model transfer (Figure 4) to achieve high accurate predictions on the succeeding domains in the transfer path, i.e., $\mathbf{D}_{i+1}$ and $\mathbf{D}_{i+2}$. Between a pair of directly transferable datasets on the transfer path, e.g., $\mathbf{D}_i$ and $\mathbf{D}_{i+1}$, *VMA* conducts a modality-aware multi-task alternative training (Section 3.4.1) and pseudo-labels $\mathbf{D}_{i+1}$ (Section 3.4.2). The trained model's shared hidden layers are reused by the succeeding pair in the path, i.e., $\mathbf{D}_{i+1}$ and $\mathbf{D}_{i+2}$, to preserve the transferable knowledge (Section 3.4.3).

3.4.1 Modality-Aware Multi-Task Alternative Training. To fairly compare the representations for two modalities and fuse data with different segmentation schemes without losing complementarity, we train a **multi-task learning** model with labeled or pseudo-labeled data. In our multi-task model, we consider recognizing activities with data from one modality as one **task**. Figure 4 shows a model with q tasks corresponding to q modalities. Each modality has its input and output layers. These modality-specific layers retain the modality's insensitivity to specific domain factor variances. The input and output layers of all modalities connect to the same hidden layers, i.e., these hidden layers are shared by q tasks [4]. This process is shown as ❶, ❷ and ❑ path in Figure 4, each path corresponds to one sensing modality. For sensing modality M_k, we denote feature vectors as $\mathbf{x}_{M_k}$, where $1 \leq k \leq q$. The model can be described as:

$$\hat{y}_{M_k} = f^O_{M_k}(f^H(f^I_{M_k}(\mathbf{x}_{M_k})), \tag{4}$$

where $\hat{y}_{M_k}$ is the prediction output, $f^I_{M_k}(\cdot)$ is the function of input layer for the sensing modality M_k, $f^H(\cdot)$ is the function of shared hidden layers. The $f^O_{M_k}(\cdot)$ is the function of output layer for the sensing modality M_k.

$$f^I_{M_k}(\mathbf{x}_{M_k}) = \phi(W^I_{M_k}\mathbf{x}_{M_k} + b^I_{M_k}), \tag{5}$$

where $W^I_{M_k} \in \mathbb{R}^{m \times n}$ is the weight matrix, m is the dimension of the next layer, n is the dimension of the input feature vector. Here $b^I_{M_k}$ is the bias term of each input layer. And ϕ is the activation function. The shared hidden layers are stacking feed-forward layers. Its function $f^H(\cdot)$ is parameterized by W^H. The shared hidden layers embed each modality's input into a comparable representation $\mathbf{z}_{M_k}$.

$$\mathbf{z}_{M_k} = f^H(f^I_{M_k}(\mathbf{x}_{M_k})) \tag{6}$$

The output layer, $f^O_{M_k}(\cdot)$, employs a softmax σ as the activation function

$$\sigma(\mathbf{z})_i = \frac{e^{\mathbf{z}_i}}{\sum_{j=1}^{J} e^{\mathbf{z}_j}}, \tag{7}$$

where $\mathbf{z}$ is the input vector to the softmax function, and i is the ith class in the total number of J classes. Intrinsically, the softmax function assigns prediction confidence to each class, and the class with the highest confidence is the predicted label $\hat{y}$

$$\hat{y} = \underset{i}{\text{argmax}}\ \sigma(\mathbf{z})_i \tag{8}$$

And the prediction confidence of $\hat{y}$ is

$$P(\hat{y}) = \max_i \sigma(\mathbf{z})_i \tag{9}$$

We employ the categorical cross-entropy loss for each modality's output layers, $\mathcal{L} = -\sum_{j=1}^{J} y_j log(\sigma(\mathbf{z})_j)$, [23], where y_j is the label. We denote the loss for the modality M_k as $\mathcal{L}_{M_k}$. The parameters of shared hidden layers and modality-specific layers are updated alternatively according to the corresponding loss

$$\mathcal{L}(\mathbf{x}) = \mathcal{L}_{M_k}, \text{ if } \mathbf{x} = \mathbf{x}_{M_k} \tag{10}$$

by backpropagation [30] with a specific optimizer. We use the Adam optimizer for its strong empirical performance [16].

We use **alternative training** to train the model, where instead of updating the model parameters with all the data from one task to another, we update them with partial data from one task then move to another task. In this way, each task updates model parameters in multiple epochs, avoiding the learned knowledge from one modality being completely overridden by another modality (catastrophic forgetting) [9]. For each epoch, we randomly select a batch of data points to train the model.

For q available modalities, the multi-task learning model outputs q predictions of the input data. *VMA* conducts the **prediction fusion**. Different sensing modalities' signal often adopts their modality-specific segmentation schemes [6]. As discussed in Section 3.1, wearable sensor's signal is segmented with sliding windows of size w_{IMU}, and their prediction output is at the window level. The structural vibration sensor's signal is processed and predicted on

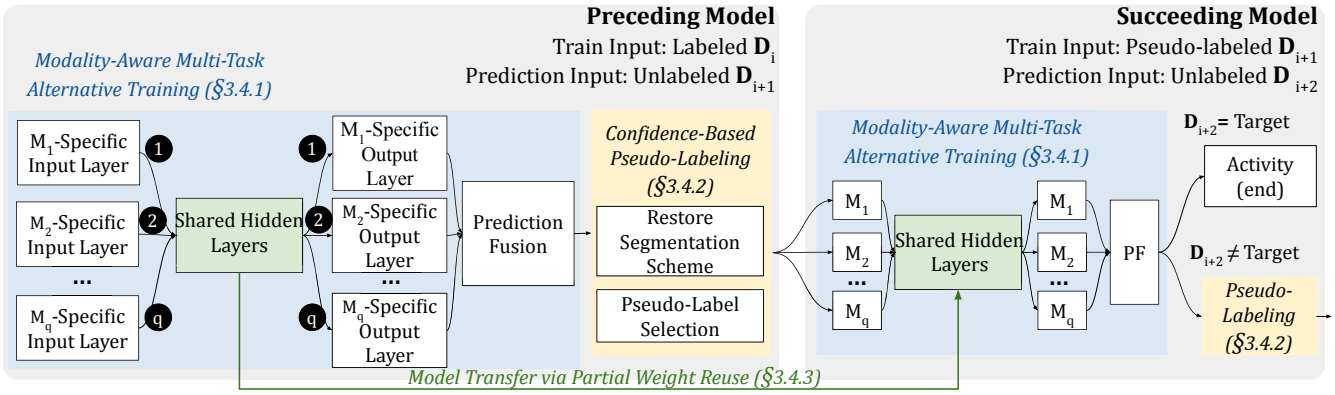

Figure 4: Modality-aware model transfer module.

event-level segments. When both segmentation schemes are used by the available modalities, *VMA* needs to align each modality's prediction into the same segmentation scheme for further fusion purposes.

For modalities adopting event-level segmentation, *VMA* first assigns the prediction and confidence to signal samples of detected events. For remaining samples without detected events, *VMA* assigns them with the non-activity label and the confidence is set as zero. Then, we apply a fusion sliding window of size w_F to the sample-level predictions. We conduct a majority vote with the sample-level predictions and output the prediction label of this window. The confidence of voted class is selected as the confidence of the window. If there are multiple confidence scores for the same voted class in the window, we select the highest one. In this paper, the size of the fusion sliding window w_F is set as the same for IMU, w_{IMU}, to simplify the modalities' prediction resolution alignment.

After aligning each modality's prediction into window level, for each fusion window, we fuse the predictions of all modalities based on their confidence scores. Since confidence scores of all modalities' predictions are generated from comparable embeddings $\mathbf{z}_{M_k}$, they are directly comparable. We choose the prediction with the highest confidence score among all modalities as the fused prediction.

3.4.2 Confidence-based Pseudo-Labeling. For an intermediate domain $\mathbf{D}_{i+1}$, the fused prediction is then used to pseudo-label the data for training the succeeding model. However, the succeeding model takes inputs in the modality-specific segmentation schemes, which may be different from the pseudo-label. Therefore, in *VMA*, we develop a restore segmentation scheme to convert different sensing modalities' pseudo-labels to their modality-specific segmentation schemes to train the succeeding model. Another challenge for training an accurate succeeding model is erroneous pseudo-labels [41]. Hence, we conduct a confidence-based selection of the fused prediction to ensure the reliability of the pseudo-labels.

Restore Segmentation Scheme. For fused predictions of the intermediate domain, e.g., $\mathbf{D}_{i+1}$ in Figure 4, *VMA* restores them back to modality-specific segmentation scheme. For modalities adopting event-level segmentation, their fused predictions are at the window level after the fusion. *VMA* converts them back to event-level to be used as pseudo-labels in the succeeding model. For events that are completely overlapped with a window (e.g., multiple events within one window), we assign the window's fused prediction as to the pseudo-label of these events. For an event that overlaps with multiple windows, we conduct a majority vote among these windows' fused predictions. The voted class label is the event's pseudo-label. In practice, we only consider those windows with more than half of their samples overlapped with the event. The confidence score of the voted class is assigned as the confidence of the event. If there are multiple confidence scores for the same voted class, we select the highest value.

Pseudo-Label Selection. To prevent negative impacts of erroneous pseudo-labels on the succeeding model [41], after restoring segmentation schemes, we conduct a pseudo-label selection. We select high-confident pseudo-labels to train the succeeding model. However, the confidence for different classes' pseudo-labels may have different value ranges, so as the same class's pseudo-labels of different modalities. To have a balanced input for the succeeding model, we apply a class-level ranked threshold on the prediction confidence within each class of each modality. For each class data of each modality, the model keeps the pseudo-labeled data with confidence in the top τ percentile.

3.4.3 Model Transfer via Partial Weight Reuse. With the selected pseudo-labeled $\mathbf{D}_{i+1}$ data, we train a succeeding model to predict the $\mathbf{D}_{i+2}$ data. However, pseudo-labels could be erroneous, which would lead to the accumulation of errors along the transfer path and directly reduce the succeeding model's accuracy. On the other hand, the preceding model is trained with labeled or less erroneous pseudo-labeled data, hence is less impacted by cumulative errors. Therefore, we reuse parameters from preceding model to constrain the cumulative error.

Reusing parameters is also challenging because the domain variance factor between $\mathbf{D}_i$ and $\mathbf{D}_{i+1}$ is often different from that between $\mathbf{D}_{i+1}$ and $\mathbf{D}_{i+2}$. Therefore, the modality that is insensitive to the domain variance factor is different between two models. The trained parameters of modality-specific layers in the preceding model may no longer be applicable to the succeeding model. As a result, the system **cannot directly reuse** the entire trained preceding model. On the other hand, parameters of shared hidden layers reflect the relationship between activity classes and modalities in

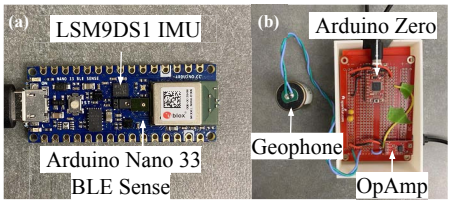

Figure 5: Sensing system hardware. (a) shows the wearable sensing node with an IMU, and (b) shows the structural vibration sensing node with a geophone sensor.

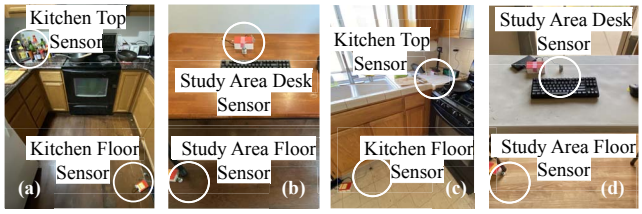

Figure 6: Structural vibration sensor experimental setup. (a) and (b) are in structure 1. (c) and (d) are in structure 2.

the preceding model. To leverage this learned relationship in the succeeding model, we reuse shared hidden layers' parameters, i.e., W^H in Section 3.4.1.

4 EVALUATION

We evaluate *VMA* with data collected in real-world experiments. We collect data with six human subjects over two residential buildings based on the IRB protocol.

4.1 Experiment Setup and Data Collection

We select 10 types of common activities of daily living (ADL) over two areas including the following activities: 1) keyboard typing, 2) using the mouse, 3) handwriting. Kitchen Area: 4) cutting food, 5) stir-fry, 6) wiping countertop, 7) sweeping floor, 8) vacuuming floor, 9) open/close drawer, and 10) idle. These activities are used to assess older adults' ability to live independently [11] and profile people's behavior [24, 38].

The sensing systems used to collect the data are shown in Figure 5. The wearable sensing device consists of an Arduino Nano 33 board and an LSM9DS1 IMU module [1] sampling at 235 Hz per axis. The structural vibration sensing device consists of an Arduino zero board with an LMV358 OpAmp and a geophone SM-24 sensor [2]. The vibration sensor is sampled at 6500 Hz. We place the on-wrist IMU sensor on the occupant and four vibration sensors in kitchen and study areas. Figure 6 shows the structural vibration sensors' placement at two structures. Structure 1 and 2 are significantly different in the layout and material. For instance, structure 1 kitchen area has a U-shape layout with a wooden floor, while the one in structure 2 has an L-shape layout with a ceramic tile floor. Also, the study desk in structure 1 is made from wood, while the one in structure 2 is made from plastic and metal. These differences would lead to different frequencies being activated even for the same excitation and therefore change the vibration data distribution between two structures [20].

We collect 12 datasets of different domains from six volunteers (three females and three males with heights ranging from 4'11"to 6'1") at two structures via semi-controlled experiments, each contains 10 trials. We define a **trial** of data as one volunteer conducting all 10 target activities at one structure. For each trial, we inform the volunteer about the type and time duration of activities to conduct. Volunteers have the freedom to decide 1) the order of activities, 2) locations within the area to conduct the activity, and 3) how the activity is done (as natural as their daily activity at home) during the data collection. The activities of each volunteer are evenly distributed in the dataset. The data is collected under the regular operation of the structure to reflect a practical ambient noise level – we do not exclude ambient noises (e.g., chronic noise from the refrigerator) in the data collection. With these datasets, we explore transfer paths in the form of $\mathbf{D}_{Source} \rightarrow \mathbf{D}_{Inter} \rightarrow \mathbf{D}_{Target}$. For a given pair of datasets with multi-factor domain variance, there are multiple decoupling solutions, i.e., the BFS may return different solutions given available intermediate domains. We generalize these **paths' properties** with a tuple describing the order of single-factor domain variance along the path. For example, $\{\mathbf{O}, \mathbf{S}\}$ represent paths, where the domain variance factor between $\mathbf{D}_{Source}$ and $\mathbf{D}_{Inter}$ is the Occupant and the domain variance factor between $\mathbf{D}_{Inter}$ and $\mathbf{D}_{Target}$ is the Structure. In our experiment setting, the transfer paths fall into two categories $\{\mathbf{O}, \mathbf{S}\}$ and $\{\mathbf{S}, \mathbf{O}\}$. We investigate 70 transfer paths under these two categories.

4.2 Modality Sensitivity Quantification

We verify that for each sensing modality, the change of directly impacting factors causes a larger domain variance than indirectly impacting factors. We quantify the similarity between two datasets of different domains via $Proxy - \mathcal{A} - distance$ (PAD), which has been used to measure the similarity between two probability distributions [10]. The lower the PAD value, the more similar the two domains [3].

We measure the PAD over pairwise datasets on both vibration sensor data and the wearable IMU data. We investigate 12 datasets of six occupants' 10 activities over two structures, and consider one of the datasets as the reference domain. Table 1 illustrates the PAD between the reference domain and the investigated domains of different structure/occupant/structure&occupant. Between the reference dataset and the dataset with structure variance, the wearable data has a lower PAD value than the vibration sensor, indicating that the wearable is less sensitive to the structure variation. Similarly, between the reference dataset and the dataset of a different occupant, we observe that the vibration sensor data has a lower PAD value. This verifies the analysis that different sensing modalities are sensitive to domain variances of different factors. In addition, when the domain variance is caused by more than one factor, we observe a higher PAD value compared to single factors'.

4.3 Evaluation Metrics

Two metrics are used to evaluate the performance of the system. First, we use the **average activity recognition accuracy** (AARA) [37] as the metric for the system evaluation. In practice, the duration

[1] https://www.st.com/en/mems-and-sensors/lsm9ds1.html
[2] https://www.sparkfun.com/products/11744

Table 1: Modality-based domain variance analysis. PAD calculated over different domain variance conditions.

Domain Variance \Modality	Vibration	Wearable
None	0	0
Structure	1.03	**0.62**
Occupant	**0.41**	1.26
Structure + Occupant	1.18	1.31

of different activities/events is different, which makes the average accuracy over absolute numbers of events activity-biased. Similar to [24], we consider the model's accuracy on each fine-grained activity equally important. Therefore, we use the average activity recognition accuracy (AARA) of the target domain (with domain variances from two physical factors).

$$AARA = \frac{1}{N} \sum_{i=1...N} Acc_i, \tag{11}$$

where N is the number of types of activities, Acc_i is the prediction accuracy of the i^{th} activity.

Second, we define **transfer success rate** (TSR) to evaluate the robustness of the system. We define the TSR as the ratio between the number of successful transfers and the total number of transfers we investigated in each evaluation experiment. Because *VMA* applies confidence-based threshold on the preceding model predictions, it is possible that the entire class of predictions is discarded. In this case, the input to the succeeding model does not contain pseudo-labeled data of all classes, and we consider it a failed transfer. On the other hand, we consider a transfer is a successful transfer when all target classes are maintained after the confidence-based threshold process. The transfer success rate is formed as:

$$TSR = \frac{N_{Success}}{N_{Total}}, \tag{12}$$

where $N_{Success}$ is the number of success model transfer in total number of N_{Total} transfers in each evaluation experiment.

For each transfer, all the labeled data from the source domain is used for training. The test is done on all the unlabeled data from the target domain. The source and target domain datasets are collected from different occupants at different structures. Given the randomized initialization for the model training, we do a 5-time repetitions to avoid outliers. We report the mean and standard deviation of the AARA over investigated transfer paths. And we calculate TSR with all repetitions over investigated transfer paths.

4.4 Implementation

4.4.1 Pre-processing. For the IMU signal, we set the sliding window size to 1.5 seconds with 0.75 seconds overlapping to capture the temporal pattern of activities while keeping a low prediction latency. For vibration signal's event detection, we apply a sliding window with size of 0.2s to ensure high segmentation precision. I.e., for activities with short impulses/pauses, this window size enables the system to capture accurate temporal segmentation of events. However, the number of events detected by vibration sensors is not comparable to the number of sliding windows of IMU signal. Therefore, To amplify the data volume, we split detected events with a length longer than 2 seconds into multiple 1-second sub-events. Then we extract frequency components from $10Hz$ to $490Hz$ as vibration signal features.

4.4.2 VMA. We train each modality-aware multi-task model with an identical structure – one input layer and one output prediction layer for each modality, and three shared hidden layers. The output dimension of each input layer is 128, and the numbers of units for each shared hidden layer is 128, 64, and 32, respectively. Dropout is applied with probability of 0.2 over hidden layers' connections [32]. The input and hidden layers adopt Exponential Linear Unit (ELU) as the activation function for its strong empirical performance and faster learning speed [5]. All modality-specific layers' parameters and shared hidden layers' parameters in the first preceding model are initialized with Xavier initialization [10]. The training batch size for each modality is 256 per batch, and the learning rate is fixed as 0.001 for both tasks. We set the confidence selection threshold τ as 40% for all evaluations unless further mentioned.

4.4.3 Learning Scheme Baseline 1: Direct Prediction. We train independent models for each modalities with labeled source domain data to directly predict the target domain (without intermediate domains). These models are feed-forward neural networks [29], which have five hidden layers with dropout [32]. The dimension of each layer, activation functions utilized, dropout rate, training batch size and learning rate are the same as those applied to *VMA* (Section 4.4.2). This baseline still employs spatial characteristics analysis (Section 3.2.1) to augment the spatial information.

4.4.4 Learning Scheme Baseline 2: Modality-aware Direct Prediction. We construct a modality-aware multi-task learning model with the same structure and training parameter setting as in *VMA*. The model is trained with labeled source domain data and then tested with unlabeled target domain data without intermediate domains.

4.4.5 VMA Oracle Mode. When there are multiple accessible transfer paths, *VMA* selects the first one coming up by the search, because there is no additional information to indicate the optimality of the paths. We further present *VMA* in the 'Oracle Mode' to compare and demonstrate the robustness of all the transfer paths. *VMA* Oracle Mode assumes that there are always sufficient datasets to generate multiple transfer paths and the best transfer path (i.e., the path that maximizes AARA) is always selected.

4.4.6 Ablation Study Baseline 1: No Confidence Threshold. In this baseline, we discard the pseudo-label selection step in the confidence-based pseudo-labeling (Section: 3.4.2). Other parts in the system are kept as same, to investigate the impact of noise and error accumulation from the pseudo-labeling.

4.4.7 Ablation Study Baseline 2: Random Transfer Path. In this baseline, we discard the domain variance-aware multi-factor decoupling (Section: 3.3). Instead, the intermediate domain is randomly assigned to investigate the contribution of the transfer path.

4.4.8 Ablation Study Baseline 3: No Spatial Modeling. Here, we aim to investigate the impact of using spatial information (Section: 3.2.1). Instead of training spatial-specific multimodal multi-task learning models, we train one model for each structure.

4.4.9 Weight Reuse Scheme Baseline 1: All Weights Reuse. To show the importance of partial weight reuse scheme, here we reuse the

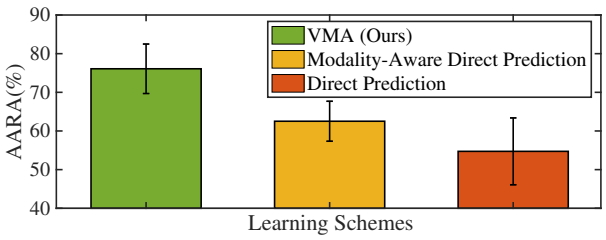

Figure 7: Comparison of different learning schemes. Compared to approaches without using intermediate domain, our *VMA* achieves an AARA of 76.1%, which is the highest among the three schemes. The modality-aware direct prediction and the direct prediction achieve an AARA of 62.5% and 54.7%.

parameters of the entire model trained in the preceding model as the initialization for the succeeding model.

4.4.10 Weight Reuse Scheme Baseline 2: Modality Independence. Here, we train two independent models for two modalities in the preceding model to pseudo-label the intermediate domain. Then each modality reuses its entire model in the succeeding model by initializing the model with weights from the preceding model.

4.4.11 Weight Reuse Scheme Baseline 3: No Weight Reuse. To evaluate the contribution of the shared hidden layers trained in the preceding model, we implement this baseline without transferring any learned model. Instead, we randomly initialize when train the succeeding model with selected pseudo-labeled data.

4.5 Learning Scheme Analysis

VMA ensures high prediction accuracy over high domain variance because it utilizes 1) selected intermediate domain (transfer path) with 2) modality-aware multi-task alternative training and 3) confidence-based pseudo-labeling.

4.5.1 Comparison of Learning Schemes. We consider direct prediction approaches as learning scheme baselines (Section 4.4.3, 4.4.4). Since the two learning scheme baselines do not adopt the confidence-based threshold mechanism, we compare only AARA without TSR here. Figure 7 shows that our *VMA* achieved a mean AARA of 76.1% with a standard deviation of 6.4%. The direct prediction without modality-aware design only achieves a mean AARA of 54.7% with a standard deviation of 8.7% due to the high domain variance between training and testing data. The direct prediction with modality-aware multi-task learning achieves a mean AARA of 62.5% with a standard deviation of 5.2%.

To verify that our modality-aware design effectively leverages the sensitivity difference of sensing modalities over different single-factor domain variances, we look into details of one example transfer path $\mathbf{D}_{Source} = [O_1, S_1] \rightarrow \mathbf{D}_{Inter} = [O_2, S_1] \rightarrow \mathbf{D}_{Target} = [O_2, S_2]$, where O_1 and O_2 are occupant ID, S_1 and S_2 are structure ID. The transfer path decouples the multi-factor domain variance between the source and target domains into two single-factor domain variances with different occupant and structure factors. Figure 8 (a) and (b) shows the prediction confidence of the modality-aware multi-task learning model on $\mathbf{D}_{Source} \rightarrow \mathbf{D}_{Inter}$ and $\mathbf{D}_{Inter} \rightarrow \mathbf{D}_{Target}$, respectively. (a) shows that the predictions on the vibration data have higher and more consistent confidences compared to the IMU data. It results in a 16% higher AARA compared to the IMU's. This is because the single-factor domain variance between these two domains is the occupant, which the wearable IMU is sensitive to. On the other hand, (b) shows an inverted trend, where the predictions of the IMU data have higher and more consistent confidences compared to the vibration data. Because the single-factor domain variance between these two domains is the structure, which the vibration sensing is sensitive to. This indicates that our modality-aware multitask learning scheme effectively models the modality sensitivity/robustness to different single-factor domain variances.

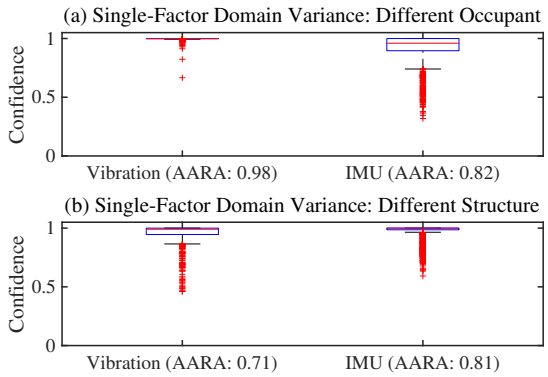

Figure 8: Prediction confidence of an example transfer path $\mathbf{D}_{Source} = [O_1, S_1] \rightarrow \mathbf{D}_{Inter} = [O_2, S_1] \rightarrow \mathbf{D}_{Target} = [O_2, S_2]$. When predicting the intermediate domain, vibration is less sensitive to the domain variance (different occupants). Its prediction accuracy and confidence are higher than those of the IMU. While when predicting the target domain, IMU is less sensitive to the domain variance (different structures) and achieves higher prediction accuracy and confidence.

4.5.2 Ablation Study. We conduct an ablation study on the design components of *VMA* with *VMA* Oracle Mode and three baselines (Section 4.4.5, 4.4.6, 4.4.7, and 4.4.8), and demonstrate their performance in AARA and TSR in Figure 9. To highlight the robustness of transfer paths, we compare *VMA* to *VMA* Oracle Mode and demonstrate the performance difference between first searched paths and optimal paths is negligible (2.8%). To demonstrate the importance of confidence-based pseudo-labeling, we compare *VMA* to the Baseline 1, where no confidence thresholding is applied. When the system does not select pseudo-labels with confidence higher than the threshold, the target domain learning accuracy drops from 76.1% to 68.5%. This is because the succeeding model has more erroneous pseudo-labels.

To understand the performance and importance of the transfer path, we investigate the *VMA* Oracle Mode and the Baseline 2. *VMA* Oracle Mode always selects the best transfer path and the Baseline 2 selects intermediate domain randomly. Comparing to the 78.9% AARA by *VMA* Oracle Mode, *VMA* achieves a comparable AARA. This is because the transfer path searching takes both modality and

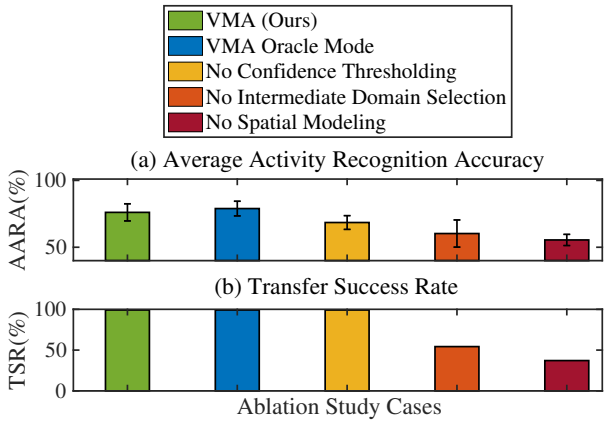

Figure 9: Ablation study. We compare *VMA* to the Oracle Mode and three baselines using AARA and TSR. *VMA* Oracle Model always select the best transfer path and therefore achieves the highest AARA of 78.9%. For each baseline, a design component is removed. *VMA* achieves AARA of 76.1%, which is slightly lower than *VMA* (< 3%). Both *VMA* and *VMA* Oracle Mode achieve a 100% TSR. The three baselines achieve AARA of 68.5%, 60.3%, and 55.5% and TSR of 100%, 54.3% and 37.1%, respectively.

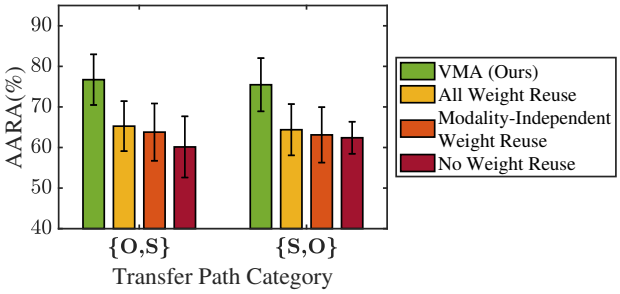

Figure 10: Model transfer schemes study. We compare *VMA* to three baselines. *VMA* adopts partial weight reuse and achieves the highest AARA. In addition, *VMA* demonstrates robustness over different categories of transfer paths.

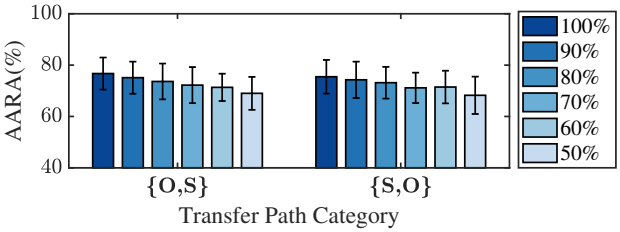

Figure 11: Target domain prediction AARA using different amounts of labeled data from the source domain.

factor into consideration. The searched transfer path is reliable for the model transfer. We observe that in Baseline 2 both the AARA and the TSR reduce, where the AARA reduced to 60.3% and the TSR is only 54.3%. Because the significant domain variance between the source and intermediate domain leads to more erroneous pseudo-labels with low confidence.

To study the importance of using spatial information (Section 3.2.1), we compare our approach to the Baseline 3, where the system does not leverage spatial modeling. The mean AARA reduces to 55.5%. Due to the structure differences (e.g., layout, material, etc.), transferring knowledge directly between different buildings is more challenging than transferring knowledge of designated areas between different buildings (e.g., kitchens of the two buildings). As a result, utilizing spatial characteristics customizes the transfer for different areas within the building to achieve a higher accuracy.

4.5.3 Comparison of Model Transfer Schemes. To demonstrate advantages of modality-aware model transfer design and its robustness over different transfer path categories, we compare it to three baselines of model transfer schemes as listed in Section 4.4.9, 4.4.11, and 4.4.10. We plot the model transfer results in Figure 10.

VMA achieves mean AARA of 76.7% for paths in the category $\{\mathbf{O}, \mathbf{S}\}$ (paths' property defined in Section 4.1) and 75.5% for for paths in $\{\mathbf{S}, \mathbf{O}\}$. The partial weight reuse allows the system to capture the modality-invariant knowledge via the shared hidden layers. While the modality-specific layers enables the system to leverage each modality's robustness in handling specific domain variance. As a result, when all the weights are reused (Section 4.4.9), the succeeding model loses its modality-specific advantage, and the mean AARA reduces to 65.3% and 64.4% for transfer paths in categories $\{\mathbf{O}, \mathbf{S}\}$ and $\{\mathbf{S}, \mathbf{O}\}$, respectively.

The shared hidden layer is important for projecting each modality input into the comparable latent space. If the models are trained in a modality-independent way (Section 4.4.10), even the weights for each modality model is reused, the modality inputs are not projected to the same latent space, which makes their prediction confidences not as comparable as those from our *VMA*. This leads to a decrease of accuracy to 63.8% and 63.1% respectively for the $\{\mathbf{O}, \mathbf{S}\}$ and $\{\mathbf{S}, \mathbf{O}\}$ transfer paths.

If we do not reuse the weights at all (Section 4.4.11), the mean AARA drops to 60.1% for paths in the category $\{\mathbf{O}, \mathbf{S}\}$ and 62.4% for paths in the category $\{\mathbf{S}, \mathbf{O}\}$. This indicates that the partially reuse weights of the shared hidden layer not only retains the shared modality-invariant information, it also allows the model to rely on the modality who is less sensitive to the domain variance for an accurate prediction.

4.6 System Parameter Analysis

We further show the system's performance with different parameter settings, including the amount of labeled data in the source domain (Section 4.6.1), the confidence-based pseudo-labeling threshold τ (Section 4.6.2), and the transfer path length (Section 4.6.3).

4.6.1 Amount of Labeled Data. Here we investigate the impact of the amount of labeled data from the source domain. Figure 11 shows the AARA of target domain prediction with different amounts of labeled source domain data. The amounts range from 100% to 50% with an interval of 10%. We observe that for paths in both categories,

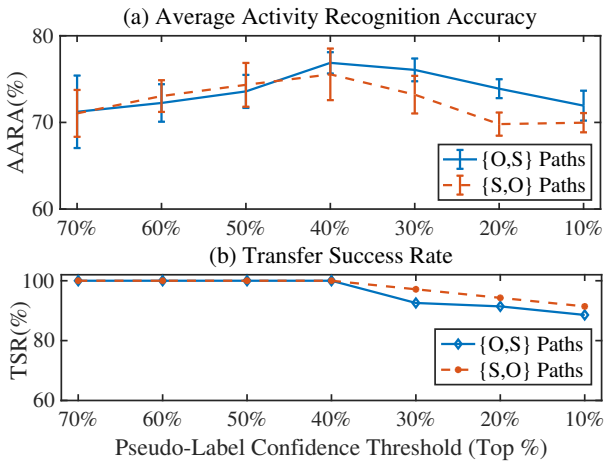

Figure 12: Target domain prediction AARA and TSR with different pseudo-label confidence thresholds.

VMA shows similar trends. For paths in the category $\{\mathbf{O}, \mathbf{S}\}$, the AARA of the target domain prediction decreases from 76.7% to 69.0%. For paths in the category $\{\mathbf{S}, \mathbf{O}\}$, their AARA varies from 75.5% to 68.2%. The results indicate that *VMA* is robust to amounts of labeled source domain data – the reduction of AARA is less than 8% given the 50% reduction in the amount of labeled data.

4.6.2 Pseudo-Label Confidence Threshold. *VMA* controls the 'quality' of pseudo-labels by selecting predictions of high confidence via class-level ranked thresholding. Therefore, the threshold value τ directly impacts the pseudo-labels reliability. We explore five different levels of threshold values and depict the AARA and TSR over two path categories $\{\mathbf{O}, \mathbf{S}\}$ and $\{\mathbf{S}, \mathbf{O}\}$ in Figure 12 (a). When the threshold value τ changes from 40% to 70%, the AARA of $\{\mathbf{O}, \mathbf{S}\}$ and $\{\mathbf{S}, \mathbf{O}\}$ decrease by 2.4% and 3.3%, respectively. This is because more erroneous pseudo-labels are included for training the succeeding model. When the threshold value τ changes from 40% to 10%, the AARA of $\{\mathbf{O}, \mathbf{S}\}$ varies from 76.7% to 71.9%, and the TSR drops from 100% to 88.5%. Similarly, for $\{\mathbf{S}, \mathbf{O}\}$, the AARA varies from 75.5% to 70.0%, with the TSR dropping from 100% to 91.4%. With the increase of the confidence threshold, the amount of the pseudo-labeled data decreases, resulting in less training data for succeeding model. In this way, the AARA of the target domain prediction decreases. In the cases where some classes have a limited number of predictions, this increase of the threshold also may cause miss class in the pseudo-label. As a result, we also observe a decrease in the TSR.

4.6.3 Case Study: Intermediate Domain Availability and Transfer Path Length. The availability of the intermediate domains directly impacts the transfer paths' length. With the increase of the transfer path length, the negative impact of erroneous pseudo-labels accumulates at each domain on the transfer path. We investigate the robustness of transfer paths generated by *VMA* with a case study by including an additional dataset of O_3 in structure S_3. We compare two paths of different lengths between the source domain $[O_1, S_1]$ and the target domain $[O_3, S_3]$. When there are limited

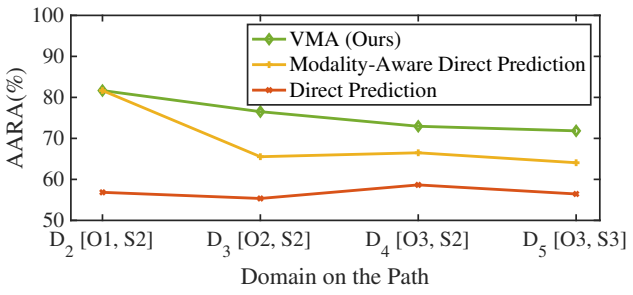

Figure 13: AARA of prediction on domains along the transfer path. Comparison between *VMA* and two baselines 1) modality-aware direct prediction and 2) direct prediction. The labeled source domain $\mathbf{D}_{Source} = [O_1, S_1]$.

available datasets from S_1, *VMA* finds a transfer path with a length of five relying on multiple S_2 datasets $\mathbf{D}_{Source} = [O_1, S_1] \rightarrow \mathbf{D}_2 = [O_1, S_2] \rightarrow \mathbf{D}_3 = [O_2, S_2] \rightarrow \mathbf{D}_4 = [O_3, S_2] \rightarrow \mathbf{D}_5 = [O_3, S_3]$. We depict the AARA of each domain in Figure 13. We observe that *VMA* is robust with domains on the transfer path, and achieves AARA of 81.7%, 76.5%, 73.0%, and 71.8% for the investigated domains, respectively. We adopt the two learning scheme baselines, as introduced in Section 4.4.3 and 4.4.4. The direct prediction baseline achieves AARA of 56.8%, 55.3%, 58.7%, and 56.4% over these investigated domains, which shows 14.3% to 24.9% lower accuracy than our *VMA*. The modality-aware direct prediction shows a comparable performance at $\mathbf{D}_2$, but yields 10.7% more error at $\mathbf{D}_5$. We compare this result with a shorter transfer path when more datasets, e.g., $[O_3, S_1]$, from S_1 are available $\mathbf{D}_{Source} = [O_1, S_1] \rightarrow \mathbf{D}_2 = [O_3, S_1] \rightarrow \mathbf{D}_3 = [O_3, S_3]$. The shorter path achieves a slightly higher AARA of 73.4% than the longer path. This higher result comes from less error accumulation.

5 RELATED WORK

5.1 Occupant Activity Recognition

There are various sensing modalities have been explored for occupant activity recognition, including device-free (e.g., RF [37], camera [17], and vibration [13]) and wearable sensing[1]. Device-free sensing usually captures human-induced signal in the space. For example, human body's interference to the WiFi signals. These systems often focus on coarse-grained activities only, due to their limited temporal resolution. On the other hand, wearable sensors continuously captures human body's motion, hence they can monitor fine-grained activities. However, due to the lack of spatial information, these systems are often limited in recognizing fine-grained motions of specific contexts, e.g., kitchen activities [21]. Comparing to these prior works, our system combines infrastructural and wearable sensing to achieve fine-grained full-home activity recognition.

5.2 Multimodal Data Fusion

Multimodal sensing systems often leverage the complementary information provided by different sensing modalities to achieve accurate inference. Various techniques to fuse the information from different modalities have been explored to make the best use of these complementary characteristics. The fusion is generally performed

at two levels: *early fusion* and *late fusion* [12]. For the early fusion, the features extracted from input data are first combined and then sent as input to a model. Common early fusion approaches include 1) explicitly concatenating feature vectors from different modalities [13] and 2) deep learning-based approaches [25], in which the learning model fuses different modalities' inputs implicitly.

On the other hand, for late fusion approaches, each modality first conducts prediction independently, then the predictions are combined using a fusion strategy, such as Bayesian inference-based weighted fusion [8]. In our study, we adopt both early and late fusion techniques to better leverage the complimentary properties between multiple modalities, yet keep each modality's insensitivity on specific domain variance.

5.3 Model Transfer

The model transfer has been successfully applied in many applications such as natural language processing[18] and computer vision[36] to ensure high learning accuracy with domain shift. Prior works have investigated both shallow machine learning [15, 35] and deep learning [28, 31] for model transfer. Shallow models often focus on leveraging data's statistical property to design the model transfer algorithm, therefore, they can work with limited amount of data. On the other hand, deep learning-based models often relay on models' feature extraction capability, e.g., latent space embedding. However, large amount of data are required to train these deep learning models. Those prior works have shown the effectiveness of applying model transfer in occupant activity recognition, however, these approaches are focusing on only single-factor of domain variance, e.g., different human subject, different spatial deployment. We focus on leveraging multiple sensing modalities and their sensitivity properties in handling different single-factor domain variances and present a framework conducts modality-aware model transfer.

6 DISCUSSION

Scalability of The Framework. *VMA* is designed as a model transfer framework and can be adjusted given the number of target classes, sensing modalities, and available computational resources.

In this work, we select 10 classes of fine-grained activities in three categories (studying, cooking, and housekeeping) to represent daily activities. The difficulty of activity recognition may increase when the number of targeting prediction classes increases. *VMA* has the flexibility to replace the activity recognition model and fit into the problem with more activity classes. In addition, since the framework takes spatial information into account, we believe scaling up to more activities over different functioning areas would have a limited impact on the model accuracy.

In addition, *VMA* is flexible with sensing modalities. We select structure vibration and wearable IMU in this study because of their complementary properties in activity recognition [13]. More sensing modalities can be included and these modalities would bring more transfer path options. For example, a camera can provide information when the ambient vibration noise is loud (e.g., robot vacuum cleaner passing by,) while the vibration sensor can provide information when lights are off.

Besides, given the available computation resource (e.g., edge devices, cloud cluster), the model complexity can be adjusted to fit into the resource. The model in this study adopts the ELU as the activation function to skip the computational complexity from batch normalization. This reduces computational load for edge device deployment. The model complexity can also be increased with a deeper and bigger network design to be deployed in clouds for more complex activity recognition tasks. In the future, we will investigate the scalability of framework given different activities classes, sensing modalities and available computation resources.

Robustness to Overlapping Signals. *VMA* is designed to conduct model transfer on activity-induced sensory signals. However, the signal of interest (SoI) can be noisy due to the overlapping of multiple person's activities or strong ambient noise. The extraction of SoI impacts the IoT sensing system performance.

In this work, we assume that there are not any forms of activity overlapping from multiple people. When there are multiple people within the same area, their activities may lead to overlapping structural vibration signals. In addition, one person may conduct multiple activities simultaneously, e.g., walking while talking or walking while cleaning. This may lead to multiple labels for the same period of time, and makes it difficult for wearables to capture. Strong ambient noises (e.g., machinery, appliance, outdoor traffics, nearby construction) would also introduce non-activity vibration events overlapping with occupant activities signals, hence negatively affecting the model transfer.

The key to improve the framework robustness to overlapping is isolating SoI from the raw data streams. Wearable sensors monitor each user independently, without being interfered with by other person's activity or ambient noise sources. This characteristic can be leveraged to assist co-located infrastructural sensors in splitting overlapped signals and identifying vibration sources. We plan to investigate splitting overlapped signals by identifying the vibration source and further improving the framework's robustness.

Domains with Different Activities and/or Bias. In this work, we focus on model transfer over the same set of activities with an even distribution. However, in real-world systems, different datasets may contain a different set of activities. For example, there is only a subset of target activities are observed in different domains. Also, the class distribution may be biased, some classes may have more data than others. In the future, we plan to explore these challenges. We plan to incorporate techniques like meta-learning and zero-shot learning, which have shown promising capability of learning new information without prior knowledge, with our model transfer framework. Also, we will investigate deep causal learning on the bias reduction for our model transfer framework.

7 CONCLUSION

In this paper, we present *VMA*, a model transfer framework for multimodal datasets with multi-factor domain variance. *VMA* first characterizes impact factors of the datasets' domain variance. For datasets of multi-factor domain variance, *VMA* decouples it to a transfer path of multiple single-factor domain variances. Next, *VMA* conducts modality-aware multi-task learning on pairs of domains along the transfer path till the target domain is predicted. We apply

VMA to the fine-grained activity recognition application with a multimodal IoT sensing system of structural vibration and wearable IMU. We conduct real-world experiments to evaluate the proposed framework and algorithm over multiple residential building structures and multiple occupants. *VMA* achieves a model transfer accuracy up to 76.1% on the target domain with multi-factor domain variance, which is 1.9× and 1.6× error reduction compared to baseline learning schemes.

ACKNOWLEDGMENT

We sincerely thank the anonymous shepherd and reviewers for their constructive suggestions. This research was supported by a 2020 Seed Fund Award from CITRIS and the Banatao Institute at the University of California.

REFERENCES

[1] Ali Akbari and Roozbeh Jafari. 2019. Transferring activity recognition models for new wearable sensors with deep generative domain adaptation. In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*. 85–96.

[2] Michael F Ashby. 2012. *Materials and the environment: eco-informed material choice*. Elsevier.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. 2007. Analysis of representations for domain adaptation. *Advances in neural information processing systems* 19 (2007), 137.

[4] Rich Caruana. 1997. Multitask learning. *Machine learning* 28, 1 (1997), 41–75.

[5] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289* (2015).

[6] L Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. 2020. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition* 108 (2020), 107561.

[7] Jonathon Fagert, Mostafa Mirshekari, Shijia Pan, Linda Lowes, Megan Iammarino, Pei Zhang, and Hae Young Noh. 2021. Structure-and Sampling-Adaptive Gait Balance Symmetry Estimation Using Footstep-Induced Structural Floor Vibrations. *Journal of Engineering Mechanics* 147, 2 (2021), 04020151.

[8] João Falcão, Carlos Ruiz, Shijia Pan, Hae Young Noh, and Pei Zhang. 2020. FAIM: Vision and Weight Sensing Fusion Framework for Autonomous Inventory Monitoring in Convenience Stores. *Frontiers in Built Environment* 6 (2020), 175.

[9] Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences* 3, 4 (1999), 128–135.

[10] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 249–256.

[11] Carla Graf. 2008. The Lawton instrumental activities of daily living scale. *AJN The American Journal of Nursing* 108, 4 (2008), 52–62.

[12] David L Hall and James Llinas. 1997. An introduction to multisensor data fusion. *Proc. IEEE* 85, 1 (1997), 6–23.

[13] Zhizhang Hu, Tong Yu, Yue Zhang, and Shijia Pan. 2020. Fine-grained activities recognition with coarse-grained labeled multi-modal data. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 644–649.

[14] Marco Iosa, Pietro Picerno, Stefano Paolucci, and Giovanni Morone. 2016. Wearable inertial sensors for human movement analysis. *Expert review of medical devices* 13, 7 (2016), 641–659.

[15] Md Abdullah Al Hafiz Khan and Nirmalya Roy. 2017. Transact: Transfer learning enabled activity recognition. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 545–550.

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Jinna Lei, Xiaofeng Ren, and Dieter Fox. 2012. Fine-grained kitchen activity recognition using rgb-d. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 208–211.

[18] Qi Li. 2012. Literature survey: domain adaptation algorithms for natural language processing. *Department of Computer Science The Graduate Center, The City University of New York* (2012), 8–10.

[19] Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. 2019. AttnSense: Multi-level Attention Mechanism For Multimodal Human Activity Recognition.. In *IJCAI*. 3109–3115.

[20] Mostafa Mirshekari, Jonathon Fagert, Shijia Pan, Pei Zhang, and Hae Young Noh. 2020. Step-level occupant detection across different structures through footstep-induced floor vibration using model transfer. *Journal of Engineering Mechanics* 146, 3 (2020), 04019137.

[21] Yasser Mohammad, Kazunori Matsumoto, and Keiichiro Hoashi. 2017. A dataset for activity recognition in an unmodified kitchen using smart-watch accelerometers. In *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*. 63–68.

[22] 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*. 158–165.

[23] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

[24] 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*. 149–158.

[25] Fan Qi, Xiaoshan Yang, and Changsheng Xu. 2018. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM international conference on Multimedia*. 429–437.

[26] Hangwei Qian, Sinno Jialin Pan, Chunyan Miao, H Qian, SJ Pan, and C Miao. 2021. Latent Independent Excitation for Generalizable Sensor-based Cross-Person Activity Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11921–11929.

[27] Eugene Rivin. 1999. *Stiffness and damping in mechanical design*. CRC Press.

[28] Seyed Ali Rokni, Marjan Nourollahi, and Hassan Ghasemzadeh. 2018. Personalized human activity recognition using convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[29] Frank Rosenblatt. 1961. *Principles of neurodynamics. perceptrons and the theory of brain mechanisms*. Technical Report. Cornell Aeronautical Lab Inc Buffalo NY.

[30] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.

[31] Andrea Rosales Sanabria and Juan Ye. 2020. Unsupervised domain adaptation for activity recognition across heterogeneous datasets. *Pervasive and Mobile Computing* 64 (2020), 101147.

[32] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[33] Seth Stein and Michael Wysession. 2009. *An introduction to seismology, earthquakes, and earth structure*. John Wiley & Sons.

[34] Yaqin Tao, Huosheng Hu, and Huiyu Zhou. 2007. Integration of vision and inertial sensors for 3D arm motion tracking in home-based rehabilitation. *The International Journal of Robotics Research* 26, 6 (2007), 607–624.

[35] Jindong Wang, Yiqiang Chen, Lisha Hu, Xiaohui Peng, and S Yu Philip. 2018. Stratified transfer learning for cross-domain activity recognition. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–10.

[36] Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing* 312 (2018), 135–153.

[37] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st annual international conference on mobile computing and networking*. 65–76.

[38] Li-Wei Wu, Wei-Liang Chen, Tao-Chun Peng, Sheng-Ta Chiang, Hui-Fang Yang, Yu-Shan Sun, James Yi-Hsin Chan, and Tung-Wei Kao. 2016. All-cause mortality risk in elderly individuals with disabilities: a retrospective observational study. *BMJ open* 6, 9 (2016).

[39] Zhong Zhang and Donghong Li. 2018. Hybrid cross deep network for domain adaptation and energy saving in visual internet of things. *IEEE Internet of Things Journal* 6, 4 (2018), 6026–6033.

[40] Xiahua Zheng and James MW Brownjohn. 2001. Modeling and simulation of human-floor system under vertical vibration. In *Smart Structures and Materials 2001: Smart Structures and Integrated Systems*, Vol. 4327. International Society for Optics and Photonics, 513–520.

[41] Zhedong Zheng and Yi Yang. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision* 129, 4 (2021), 1106–1120.

[42] Zhijun Zhou, Yingtian Zhang, Xiaojing Yu, Panlong Yang, Xiang-Yang Li, Jing Zhao, and Hao Zhou. 2020. Xhar: Deep domain adaptation for human activity recognition with smart devices. In *2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.