ViTag: Online WiFi Fine Time Measurements Aided Vision-Motion Identity Association in Multi-person Environments

Bryan Bo Cao
Stony Brook University
boccao@cs.stonybrook.edu

Abrar Alali
Old Dominion University
Saudi Electronic University
aalal003@odu.edu

Hansi Liu
Rutgers University
hansiiii@winlab.rutgers.edu

Nicholas Meegan
Rutgers University
njm146@scarletmail.rutgers.edu

Marco Gruteser
Rutgers University
gruteser@winlab.rutgers.edu

Kristin Dana
Rutgers University
kristin.dana@rutgers.edu

Ashwin Ashok
Georgia State University
aashok@gsu.edu

Shubham Jain
Stony Brook University
jain@cs.stonybrook.edu

***Abstract*—In this paper, we present *ViTag* to associate user identities across multimodal data, particularly those obtained from cameras and smartphones. *ViTag* associates a sequence of vision tracker generated bounding boxes with Inertial Measurement Unit (IMU) data and Wi-Fi Fine Time Measurements (FTM) from smartphones. We formulate the problem as association by sequence to sequence (seq2seq) translation. In this two-step process, our system first performs cross-modal translation using a multimodal LSTM encoder-decoder network (*X-Translator*) that translates one modality to another, e.g. reconstructing IMU and FTM readings purely from camera bounding boxes. Second, an association module finds identity matches between camera and phone domains, where the translated modality is then matched with the observed data from the same modality. In contrast to existing works, our proposed approach can associate identities in multi-person scenarios where all users may be performing the same activity. Extensive experiments in real-world indoor and outdoor environments demonstrate that online association on camera and phone data (IMU and FTM) achieves an average Identity Precision Accuracy (IDP) of 88.39% on a 1 to 3 seconds window, outperforming the state-of-the-art Vi-Fi (82.93%). Further study on modalities within the phone domain shows the FTM can improve association performance by 12.56% on average. Finally, results from our sensitivity experiments demonstrate the robustness of *ViTag* under different noise and environment variations.**

***Index Terms*—Cross Modal, Fine Time Measurements, Inertial Tracking, Object Tracking, Association**

I. Introduction

With the plethora of sensors surrounding us, associating user identity across multiple sensing modalities can be significant in supporting multi-view learning across heterogeneous sensors. Multimodal association has the potential to lend itself to a wide range of applications that need cross-modal identification, such as localization, re-identification, and continuous tracking. With the pervasive use of cameras and smartphones, a key application scenario is the association between persons detected in camera video and sensor data captured from their smartphones, as depicted in Figure. 1. An example application includes sending alert messages to specific users' devices who have been detected on camera (even when their face may not be visible); a particular use case for this is in facilitating exposure notifications during the current COVID-19 pandemic.

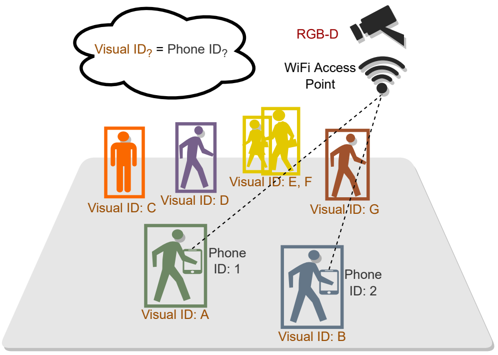


Fig. 1. Motivation: Associating visually detected subjects with corresponding phone identifiers using multimodal data.

In another scenario, to improve traffic safety, distracted pedestrians at risk can be detected through an infrastructure mounted camera and alerted by voice or vibration on their smartphones. Further, in emergencies, evacuation instructions can be sent to a person's device depending on their precise location. All these scenarios require associating the camera image with a device identifier.

To associate data across modalities, existing approaches require predefined visual features (e.g. clothes color [34] or gestures [7]), multiple IMU devices placements [28] (such as back at hip height), calibrated IMU and camera coordinates [11], limited depth changes [2], visible finger movements from the camera's field of view [18]. Few past works have focused on correlating visual and inertial data [2], [9]. However, these systems do not provide real-time association are therefore not usable in real-world scenarios. Moreover, techniques relying on hand-crafted features [2] often fail in more complex scenarios, where lighting condition varies, multiple people exit and re-enter the camera view, etc. Kwon

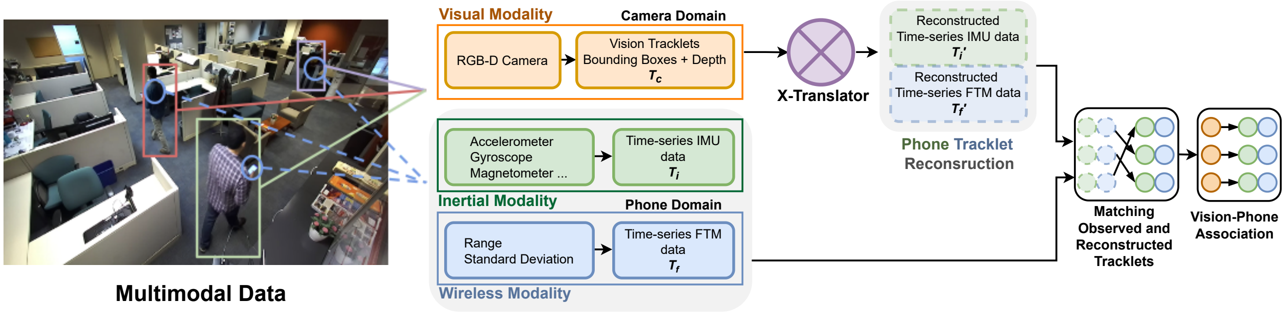


Fig. 2. *ViTag* System Overview. The system first translates data from camera domain to phone domain using the proposed model *X-Translator*, then it finds the matching between the reconstructed and observed phone data. Vision tracklets (T_c) are fed into *X-Translator* to reconstruct the corresponding phone tracklets for IMU (T_i') and FTM (T_f').

et al. [16] and Rey et al. [26] presented an automated pipeline that converts videos of human activities to inertial data for training common Human Activity Recognition (HAR) models. However, these methods focus on capturing salient features for activity recognition rather than on more stringent features for disambiguation of multiple people that may all perform the same activity (e.g., walking). In addition, prior works [23]–[25], [36] with encoder-decoder architecture that learns a joint representation among vision, inertial and especially WiFI FTM data simultaneously is under investigated. Our proposed work takes this line of research further by associating identities for subjects walking in a scene captured in multimodal data.

Approach. In this paper, we present *ViTag*, which associates data across camera and phone domains. Specifically, a vision tracker is used to generate tracklets from the camera frames. These tracklets are then matched with IMU and FTM data obtained from the smartphones. Our framework consists of a cross-modal encoder-decoder network *X-Translator*, which employs bidirectional LSTM and a joint representation between visual data from camera, and motion and WiFi data from smartphones. *X-Translator* leverages the joint representation to reconstruct or translate one modality into the other. The reconstructed data (e.g. reconstructed phone data) is then matched in real-time with the observed data from that modality (captured IMU measurements or FTM). Given the privacy questions that this approach raises, we focus on scenarios where users have opted-in to share sensor data from their phones with a camera and access point setup for such applications.

Challenges. Multimodal learning presents unique challenges due to the heterogeneity of the data. In particular, correlating camera and phone sensing data poses two significant challenges. First, each sensing modality captures data in a different coordinate space. This requires a system to be capable of transforming data from camera coordinates to a local reference frame, such as that of the IMU. Second, each modality offers varying levels of data fidelity. For example, visual sensors are less useful in low lighting conditions even when inertial data quality is not affected. Meanwhile, inertial sensor data exhibits drifting and accumulating biases over longer duration. Similarly, distance and error estimation in FTM data can be affected by multipath.

Contributions. In addressing the above-mentioned challenges, we make the following contributions:

- We design and develop *ViTag* to associate identities for subjects detected across camera (vision) and smartphone (motion and FTM) data, achieving an online association IDP of 88.39% on average.
- We propose a cross-modal encoder-decoder architecture, *X-Translator*, that learns the joint representation between camera and phone domains, and translates vision trajectories to phone (IMU+FTM) readings and vice versa.
- *ViTag* is robust to sensor noise and scene changes from indoor to outdoor multi-person environments, achieving association accuracy (IDP) of 90.21%, 87.85%, and 87.11% in *Indoor*, *Outdoor*, and *Crowded* dataset, respectively.

II. System Overview

ViTag associates identities across two domains: camera and smartphone. In the context of a camera observing humans in the scene, such as in Figure 1, the goal is to identify which device is being carried by which subject in the scene. Figure 2 shows an overview of the system. We install an RGB-D camera with a WiFi access point overlooking a large space. Users in the scene walk around with their smartphones, where each device captures accelerometer, gyroscope, and magnetometer data, while exchanging FTM messages with the access point.

Subjects are detected and tracked in the camera data using state-of-the-art trackers to generate tracklets. Each tracklet is a sequence of bounding boxes in the camera coordinate system. *ViTag* deploys a two-step process to associate multiple modalities. First, cross-modal translation, wherein our proposed *X-Translator* takes these vision tracklets as input and reconstructs the corresponding phone data, including time-series IMU readings and FTM data. To the best of our knowledge, *X-Translator* is the first network to jointly learn inertial motion, visual data, and wireless modality. Second, we use maximum bipartite matching (Hungarian Algorithm) to match the reconstructed phone data with the data received from the phone domain in real-time.

III. Cross-Modal Translation

A. Preprocessing Workflow

Camera data. We employ the StereoLabs ZED tracker for generating trajectories (referred to as *tracklets* in the rest of the paper) from an RGB-Depth camera data. Tracklets are generally short in length since subjects move out of the camera view frequently. Tracklets from camera data (T_c) are represented as a time-series sequence of bounding boxes (*BBX*). Each bounding box is represented as:

$$BBX = [x, y, d, w, h]; \quad T_c \in \mathbb{R}^{K \times 5} \tag{1}$$

where x and y are the coordinates of the bounding box centroid, d is the centroid's depth measurement, and w and h are the bounding box width and height.

Phone data. To preprocess the smartphone data, we concatenate 6 types of measurements from the time-series IMU data:

$$T_i^t = [acc; grav; lin; mag; gyro; q]; \quad T_i \in \mathbb{R}^{K \times 19} \tag{2}$$

where acc represents the 3-axis accelerometer data and $grav$ and lin are the gravitational and linear components of the accelerometer data. $gyro$ and mag represent the 3-axis gyroscope and magnetometer data. q represents the 4-axis quaternion data. An FTM measurement at time t is defined as:

$$T_f^t = [r, std]; \quad T_f \in \mathbb{R}^{K \times 2} \tag{3}$$

where r indicates the estimated range, or distance from phone to WiFi access point, while std represents the standard deviation calculated in a single RTT burst.

In the context of our work, one *modality* refers to one type of data such as bounding boxes, IMU readings, or FTM data, while *domain* refers to the source, such as camera or smartphone. Therefore, vision tracklets (T_c) are considered in the camera domain and the phone domain (T_p) consists of IMU and FTM data:

$$T_p = [T_i; T_f]; \tag{4}$$

To enable accurate real-time association, we process and match the data within limited time windows. While longer time-series windows may contain more discriminative features that can be useful for association, they add latency to the association task. To address this trade-off, we empirically choose a window size, K, to be 10 samples for all modalities in both domains. Due to slightly different sampling rates in indoor and outdoor dataset, 10 samples equals to 3 seconds in indoor environment and 1 second in outdoor environment. Therefore, association is always performed in less than 3 seconds. At each time step, the most recent K samples are used as inputs into the network.

Synchronization. Due to different sampling rates and timestamps, we need to synchronize all the modalities before feeding them to the model. We synchronize the camera and phone data using Network Time Protocol (NTP) on the devices. The sampling rate for camera frames is 30 fps, for IMU readings is 100 Hz, and 3-5 Hz for FTM. Moreover, camera (BBX) and phone (IMU, FTM) data have 16 and 13 precision timestamps. We use data from the camera domain (downsampled to 10 fps) as anchor to resample other modalities. Specifically, for each camera frame, we find the IMU and FTM readings with timestamps that are closest.

B. Network Architecture Design

The design of *X-Translator* is inspired by the self-supervised ability of autoencoders. Unlike unimodal autoencoders, *X-Translator* requires labeled correspondences between camera and phone data. The novelty of the proposed architecture lies in the application of the autoencoder approach for multiple modalities. *X-Translator* consists of three main modules: (1) an Encoder that learns the unimodal representations for each input modality, (2) a joint representation layer that learns the cross-modal latent features, and (3) a Decoder to reconstruct each modality. The architecture is depicted in Figure 3.

X-Translator consists of several instances of encoder, each learning the representation for one modality. An encoder includes a 1D convolutional layer with 32 filters, kernel size of 16 and stride of 1, followed by a ReLU activation function. Then a bidirectional LSTM layer extracts temporal features from the IMU and vision modalities, in both directions between earlier to later frames. A joint representation integrates features extracted from the unimodal data streams into a single multimodal representation by summation. Lastly, each decoder consists of two stacked bidirectional LSTM layers to extract the fused features in a hierarchical way. We use the term *codec* to refer to a pair of encoder and decoder for the same modality.

Model Loss Functions. To learn multimodal translation, we design our loss functions as follows:

- **Self-reconstruction Loss:**

$$L_{\text{self}} = \sum_{m \in M} L(D_m(E_m(X_m)), X_m) \tag{5}$$

- **Cross-modal Reconstruction Loss:**

$$L_{\text{cm}} = \sum_{m \in M} L(D_{\overline{m}}(E_m(X_m)), X_{\overline{m}}) \tag{6}$$

- **Cross-domain Reconstruction Loss:**

$$L_{\text{cd}} = L(D_p(E_c(X_c)), X_p) + L(D_c(E_p(X_p)), X_c) \tag{7}$$

- **One-to-all Reconstruction Loss:**

$$L_{\text{1-to-all}} = \sum_{m \in M} L(D_M(E_m(X_m))), X_M) \tag{8}$$

- **Fused-reconstruction Loss:**

$$L_{\text{fused}} = \sum_{m \in M} L(D_m(E_M(X_M)), X_m) \tag{9}$$

- **Multi-reconstruction Loss:**

$$L_{\text{multi}} = L(D_M(E_M(X_M)), X_M) \tag{10}$$

Our final loss function for reconstruction is the weighted sum of these losses:

$$\begin{aligned} L = &\lambda_{\text{self}} L_{\text{self}} + \lambda_{\text{cm}} L_{\text{cm}} + \lambda_{\text{cd}} L_{\text{cd}} + \\ &\lambda_{\text{1-to-all}} L_{\text{1-to-all}} + \lambda_{\text{fused}} L_{\text{fused}} + \lambda_{\text{multi}} L_{\text{multi}} \end{aligned} \tag{11}$$

where E_m and D_m are modality m's encoder and decoder, respectively. M is the set of all modalities: $m \in M, M = \{c, i, f\}$; p represents both i and f in the phone domain. $\overline{m}$

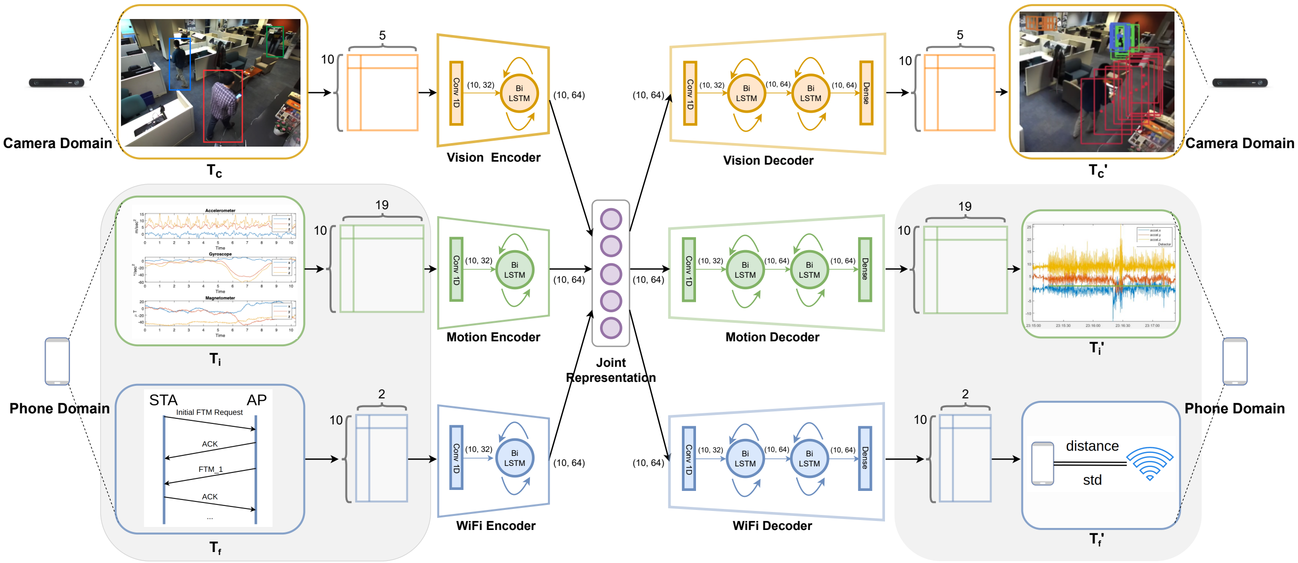


Fig. 3. *X-Translator* architecture: A bidirectional LSTM based encoder-decoder model. Encoders are used to learn unimodal representations from vision tracklets (T_c) and IMU data (T_i). A joint represention is then learned for the two modailities, implemented by element-wise summation layer. In the final layer, Decoders translate one modality to another.

denotes a different modality other than m in the L_{cm}, $L_{\text{1-to-all}}$, L_{fused} and L_{multi} loss functions. For simplicity of notation, we use E_p, D_p to refer to the encoder-decoder in the phone domain, while in real implementation there are two separate IMU and FTM encoders and decoders. Similarly, E_M or D_M denotes separate encoders or decoders for all three modalities.

The general intuition behind these loss functions is to train the network to learn a joint representation across camera and phone domains for cross-modal reconstruction while each serves its specific purpose. L_{self} optimizes the weights for reconstruction from the same modality; L_{cm} helps the network to reconstruct modality from one to another; L_{cd} forces the model to learn cross-domain reconstruction; the network is forced to reconstruct all modalities given only one modality by $L_{\text{1-to-all}}$; L_{fused} is used to learn constructing one modality when feeding all modalities, and L_{multi} loss enforces the network to reconstruct all modalities when all inputs are available. The One-to-all Reconstruction Loss forces the network to learn to reconstruct all the other modalities when only one of them is available as input. Absent data is represented with zero. We set equal weights for all losses.

During evaluation, only one modality is used as input and data in the other domain is reconstructed. The reconstructed modality is then used for association, as discussed in the following section.

C. Reconstruction Paths

Given the input modalities from two domains, there are two possible choices for reconstruction: (a) reconstructing phone data from vision tracklets, and (b) reconstructing vision tracklets (bounding box sequences) from phone sensors data. We explore both choices.

Reconstructed data is denoted with a prime ($'$) sign. For example, T_p denotes data captured from the smartpthone device; T_p' denotes phone data reconstructed using *X-Translator*.

- $T_p \longrightarrow T_c'$: Bounding box sequences are reconstructed from phone data.
- $T_c \longrightarrow T_p'$: Sequences of phone data are reconstructed from vision tracklets.

Unless otherwise specified, phone domain data includes both IMU readings and FTM.

IV. Association

X-Translator is a cross-modal translator that translates camera data into smartphone (inertial and FTM) data and vice versa. After translating one modality into another, we perform association on observed data and that reconstructed by *X-Translator*.

Association by Bipartite Matching. The association problem is formulated as finding the global minimum-weight matchings in a bipartite graph, which is also referred to as linear assignment problem. For matching T_p' and T_p, the first step is to define a distance, or dissimilarity function, and formulate the association problem in a graph setting where nodes represent modalities and edges' weights denote the distances between a possible assignment between two modalities.

We define a graph $G = (V, E)$ where V represents its nodes and E is the set of edges between the nodes. Nodes are divided into two parts. We denote V_p as the set of nodes that represents phone tracklets, and V_p' as the nodes that represent reconstructed phone tracklets. $|V_p|$ and $|V_p'|$ vary in different scenes. In the *Indoor* dataset when $|V_p| \geq |V_p'|$, i.e. the number of detected track IDs is greater than the number of reconstructed track IDs. This is because not all users are in the camera view at most times, but all phone IDs are "visible" to the system at all times. On the other hand, $|V_p| \leq |V'_p|$ in the *Outdoor* dataset because many passers-by (participants without phone data exchange) are detected in the camera view. We assign edge weights by computing the dissimilarity between every pair of nodes across $V_{p'}$ and V_p. The Hungarian algorithm [15] is used to find an optimal matching from V_p to V'_p that minimizes the total weights of the edges, which essentially maximizes the similarity of candidates between two sets of nodes. From the optimal matching we find the association between vision tracklet IDs and smartphone IDs.

Distance Function. A common dissimilarity, or distance function for association is Euclidean Distance (ED). When different modalities are translated into one common modality, ED is appropriate to measure the distance of multidimensional data, e.g. a 19 dimensional vector for IMU data, and can be applied to different reconstructed modalities without modification. With ED, the association performance can achieve around 70% to 80% in the phone reconstruction path. We use ED: Euclidean Distance $||T_m - T'_m||_2$ as the default distance function for each modality m, where T' is the reconstructed data that shares the same modality with observed data T.

We further investigate our data, especially FTM, which consists of a two dimensional vector (r_f, σ_f) where r_f is the range estimate and σ_f is the standard deviation. Since a mean and a standard deviation can define a Gaussian distribution, each FTM data point essentially can be treated as a different Gaussian distribution. To exploit FTM data points' statistical characteristics, we propose to explore Bhattarcharya Distance (BD) [5], [6], [8] over ED for the FTM modality. The intuition of Bhattarcharya Distance is to measure the separability of two distributions considering the overlap. By definition, the Bhattarcharya Distance of two distributions is

$$BD(f, f') = \frac{1}{4} \ln\left(\frac{1}{4}\left(\frac{\sigma_f^2}{\sigma_{f'}^2} + \frac{\sigma_{f'}^2}{\sigma_f^2} + 2\right)\right) + \frac{1}{4}\left(\frac{(r_f - r_{f'})^2}{\sigma_f^2 + \sigma_{f'}^2}\right) \quad (12)$$

where f is an FTM vector at one frame. Note that T_f is time-series FTM data consisting of K samples of f.

V. Experiments and Evaluation Setup

There exist several multi-modal datasets in the literature. However, these datasets are used for different tasks such as human activity recognition (HAR) based on IMU data [21], [30]. To the best of our knowledge, comprehensive datasets that include camera data, IMU sensors, and WiFi FTM measurements do not exist. Therefore we conducted an IRB approved study to collect a large-scale dataset including the aforementioned modalities, both in indoor and outdoor environments.

In this section, we describe our experimental setup, data preparation and training, and the metrics for evaluation.

A. Experiment Setup

We set up one ZED-2 stereo camera, capable of capturing RGB frames and Depth (RGB-D), and one Google Nest WiFi access point (AP) mounted next to each other on the lab room's ceiling (for indoor test environment), or mounted on a car-roof-mounted bike handle for the outdoor environment (to simulate a common pole mounting scenario for outdoor cameras) as shown in Figure 4. The camera and AP are mounted next to each other to simulate WiFi-enabled cameras that are becoming increasingly common. During the experiments, the devices collect multi-modal data from the WiFi AP, camera, and each user's smartphone. The proposed *X-Translator* model is implemented on a linux PC(Ubuntu 18.04) equipped with one NVIDIA GeForce RTX 2080 SUPER graphics card in Keras 2.4.3, Tensorflow 2.3.0 and Python 3.7.

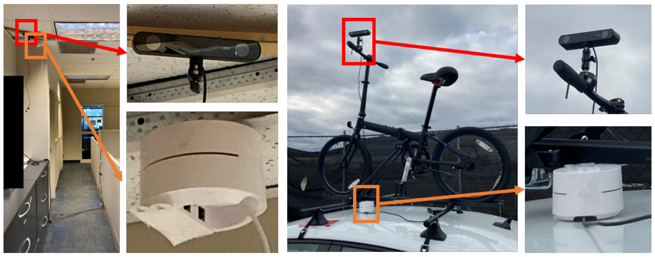
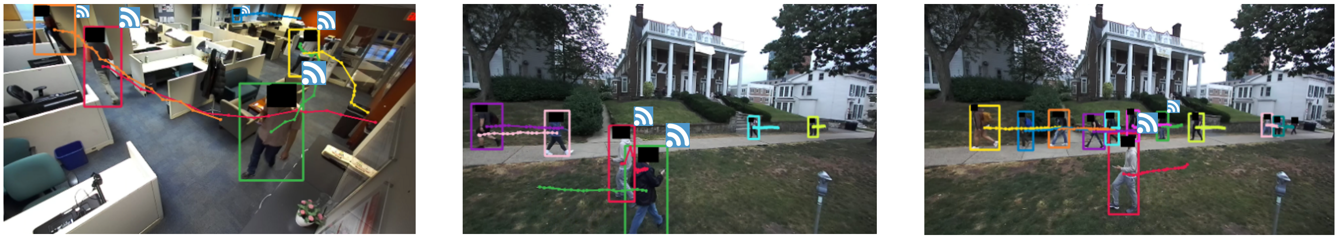


Fig. 4. Experiment Setup. A StereoLabs ZED2 camera and a Google Nest WiFi AP mounted on the ceiling in the *Indoor* setup while the camera is mounted on the handle of a roof-mounted bike for outdoor setup. This setup simulates increasingly common WiFi-enabled cameras.

B. Dataset

We present three datasets: *Indoor*, *Outdoor*, and *Crowded*. We captured 31 video *sequences* of 3 minutes duration each. On an average, we obtained 575 and 1800 frames per video sequence for *Indoor* and *Outdoor* datasets. The indoor experiments involved 5 subjects, each carrying a Google Pixel 3a smartphone, and walking at freewill and in random fashion (no path constraint was set) across the room. The crowded dataset was collected only in outdoor settings due to COVID 6-feet social distancing restrictions in indoor environments. Figure 5 shows an example of sampled frames from the labeled *Indoor*, *Outdoor*, and *Crowded* datasets. The data collection process includes, (i) FTM messages exchanged between AP and each phone at 3-5 Hz, (ii) 9-axis IMU data (acc, gyro, mag) on each user's phone at 100Hz rate, (iii) activity logs on the AP, and (iv) RGB-D camera footage captured at 30FPS at 720p resolution (1280 x 720). The experiments were conducted over multiple days following COVID protocols and restrictions. Participants (5 for indoor and 2 for outdoor) were not restricted in how they carried the phones. Passer-by pedestrians' phones did not communicate with the access point. IMU data were collected from Google Pixel 3a smartphone devices. In addition, magnetometer and linear acceleration data were recorded, and the quaternion data were computed for the dataset. The maximum number of detected pedestrians (phone holders and passerby) at a time is 11. A participant with a phone, however, might exit and re-enter the camera's field of view due to unconstrained walking pattern and limited field of view of the camera. As a result, the number of pedestrians detected in vision modality (denoted as $|V|$) could be less than, equal to, or greater than the number of participants' phones (denoted as $|P|$) detected over the wireless channel. This change of



(a) Indoor (b) Outdoor (c) Crowded

Fig. 5. Sample frames from the *Indoor*, *Outdoor*, and *Crowded* dataset with bounding box information. Tracklets shown were computed using most recent 20 frames. Best viewed when zoomed. All subjects walk in unconstrained patterns. Pedestrians whose phones are communicating with the access point are annotated with the WiFi symbol.

cardinality in both modalities poses a challenge to the cross modal association.

C. Data Preparation and Training

Data preparation. Mounting cameras in the same position but in different sessions unavoidably leads to various camera perspectives. We ensured that the video frames over multiple days were aligned in space (area where the users were walking) by applying the homography matrix to adjust the camera frames' perspective in the *Indoor* dataset, based on SIFT [19] features in a common space between views. As discussed in Section III, a vision tracklet (T_c) is represented as a sequence of bounding boxes, where each bounding box is represented in 5 dimensions (T_c^t), as shown in Equation 1. IMU data is represented as T_i, a sequence of concatenated sensor data, as shown in Equation 2. At each time t, the concatenated vector T_i^t contains 19 dimensional data. In addition to IMU, phone data includes FTM vectors.

Training. We train *X-Translator* using paired camera-phone data. For ground truth visual data, we manually annotate each frame in the dataset with bounding boxes to construct tracklets. We also use the ZED tracker for obtaining trajectories from vision data, as a secondary approach to obtain tracklets for our analysis. In both cases, the trajectories are labeled with the phone ID (unique pseudo IDs provided by our data collection application on the phone)—providing us the ground truth for association. Passers-by (without phone data connection to our network) are labeled as *Others*. Adam optimizer and Mean squared error (MSE) are used to train *X-Translator* with learning rate 0.001 and batch size 32. [1] Mean squared error (MSE) is applied to each loss function in Equation 11.

D. Evaluation Metrics

The primary metric for evaluating *ViTag* is IDentification Precision [27] or IDP, defined as: $IDP = \frac{IDTP}{IDTP+IDFP}$ where $IDTP$ and $IDFP$ are IDentification True Positives and IDentification False Positives, respectively. IDP is calculated for each association window of 10 samples (K=10). The system is evaluated in an online mode, wherein the K most recent samples are processed to determine association. Our system is evaluated via Leave-One-Out Cross Validation (LOOCV) for *Indoor* and *Outdoor* datasets. Each dataset consists of 15 sequences. To evaluate our system in a more challenging crowded scenario, we specifically hold out one sequence with maximum 11 subjects in one frame as the $Crowded$ test set to evaluate the models trained in the $Outdoor$ dataset.

E. Baselines

We compare *ViTag* with two baselines. The first is a handcrafted association technique that relies on pedestrian dead reckoning (PDR) [33] and procrustes analysis (PA) [10], [14]. As a second baseline, the state-of-the-art Vi-Fi [17] is used as the deep learning baseline.

Handcrafted Baseline. We compute 3D trajectories from camera and phone data and match them based on shape similarity, measured using PA. Trajectories from the inertial sensor data are computed using PDR, wherein the heading ϕ, computed from the accelerometer and magnetometer, is combined with average step length for adults ($l = 0.8m$) to determine the next phone position. A 3D position in the phone domain is defined as $\hat{T}_p^t = (x_p^t, y_p^t, r_p^t)$ where r_t is FTM range. Vision trajectories are computed from bounding box centroid coordinates and depth where a 3D point is defined as $\hat{T}_c^t = (x_c^t, y_c^t, d_c^t)$. We normalize (x_p^t, y_p^t) such that it is in the same scale of (x_c^t, y_c^t).

Then PA is used in the next step for association. PA measures shape similarity between two matrices, where the optimal matrix transformation (including scaling/dilation, rotations and reflections) from one to the other is applied such that the sum of squared differences between two is minimized. We follow the bipartite matching method described in Section IV to associate $\hat{T}_p$ and $\hat{T}_c$, using the dissimilarity score between their shapes (obtained from PA) as the edge weights to find the association. We believe that this is an appropriate handcrafted baseline since it employs commonly used techniques wherein shape similarity can deal with the difference in coordinate systems between phone (local frame of reference) and camera (image plane). We use PDR+PA to refer to this method.

Deep Learning Baseline. Vi-Fi [17] employs a two-stream LSTM based encoder for camera and phone modalities (IMU and FTM) separately, followed by feature ensemble and dimension reduction layers. In the last step, an affinity matrix layer is learned to predict an association decision between

[1] Code is available at https://github.com/bryanbocao/vitag. Dataset can be downloaded at https://sites.google.com/winlab.rutgers.edu/vi-fidataset/home.

two modalities. Different from Vi-Fi [17], the *X-Translator* in *ViTag* uses a joint representation and separate modality decoders without an affinity matrix layer.

VI. Evaluation

A. Overall Performance

Evaluation of Association. *ViTag*'s performance as compared with the baselines is summarized in Table I. Overall, our system achieves the highest association performance. Specifically, *ViTag* achieves an average IDP of **88.39%** in all datasets, higher than PA+PDR of 38.41% by 49.98% and Vi-Fi of 82.93% by 5.46%, respectively.

Method	PDR+PA [10], [14], [33]	Vi-Fi [17]	**ViTag (Ours)**
Avg. IDP	38.41%	82.93%	**88.39%**

TABLE I
SUMMARY OF ONLINE ASSOCIATION PERFORMANCE IN ALL DATASETS

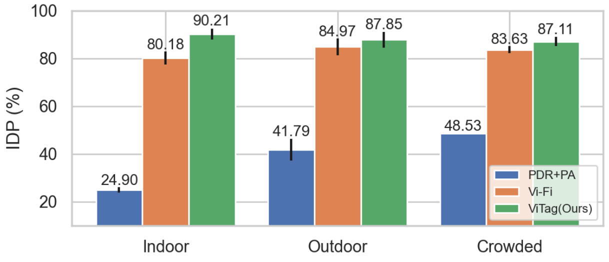
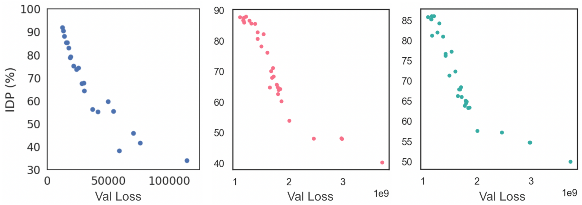


Fig. 6. *ViTag* LOOCV online association performance in green compared with handcrafted PDR+PA and SOTA deep learning (Vi-Fi) approaches in blue and orange, respectively; *ViTag* outperforms both baselines.

ViTag outperforms both baselines, achieving an average IDP of 90.21%, 87.85%, and 87.11% in the *Indoor*, *Outdoor* and *Crowded* test set, respectively, as shown in Figure 6. Note that the association results are performed on a 10-sample sliding window with 90% overlap, which corresponds to ~3 seconds in *Indoor* and 1 second in the *Outdoor* dataset.

Effect of Reconstruction. *ViTag* performs cross-modal reconstruction followed by association. We aim to understand how the reconstruction ability of *X-Translator* affects association.



(a) *Indoor* Seq 2 (b) *Outdoor* Seq 7 (c) *Crowded* Seq 7

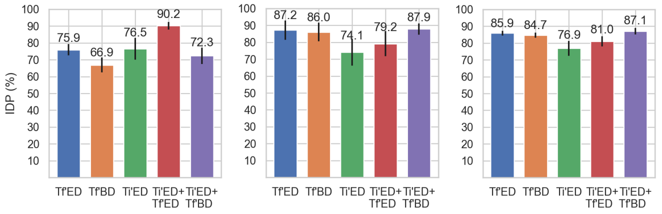
Fig. 7. Effect of validation loss on IDP for models tested in Seq 2 in *Indoor*, and Seq 7 in *Outdoor* and *Crowded* datasets by LOOCV. X-axis represents validation loss. Observe better reconstruction quality (low validation loss) improves association performance (high IDP at the top left corner).

To this end, we analyze a set of models with different weights during training. A model is denoted by a test sequence (Seq) number by LOOCV, specifically Seq 2 and 7 shown in Figure 7. Weights are saved separately during training. Validation loss of reconstruction measures cross-modal reconstruction quality. Weights with large validation loss represent models in the earlier stage of training. The relationship between validation loss and IDP for models tested in two sequences (Seq 2 in the *Indoor*, Seq 7 in the *Outdoor* and *Crowded* datasets) is shown in Figure 7. Observe that a higher cross-modal reconstruction quality (lower validation loss) tends to result in higher IDP.

B. Micro-benchmarks

In this subsection, we analyze *ViTag*'s sensitivity to various factors of system settings.

1) Effect of FTM distance function: We experiment how different distance functions impact association. In addition to Euclidean Distance (ED), we also explore Bhattacharyya Distance function (BD). $T_i'ED$ denotes association performance based on camera-IMU modalities only using ED, by ablating the Wi-Fi modality.



(a) Indoor (b) Outdoor (c) Crowded

Fig. 8. Impact of different distance functions on association performance.

There are several key observations. First, combining FTM with IMU data improves the association performance for all datasets shown in Figure 8. Second, $T_i'ED{+}T_f'ED$ works the best (IDP 90.21%) in the *Indoor* dataset, but decreases when we employ BD for FTM. However, we see the reverse trend in the *Outdoor* and *Crowded* datasets that $T_i'ED{+}T_f'BD$ results in the highest IDPs of (87.85% and 87.11%), respectively. One reason behind this could be the different data distributions in *Indoor* and *Outdoor* environments that consist of various multipath profiles, leading to variations in FTM ranging performances. Specifically, *Indoor* dataset features high-multipath environment in the lab consisting of doors, walls and so forth, while more open space exists in the *Outdoor* dataset, resulting in diverse FTM variance [12]. Experiments demonstrate the advantage of BD function over ED for FTM via performance improvement from 79.18% to 87.85% (*Outdoor*) and 81.01% to 87.11% (*Crowded*).

2) Effect of IMU Noise: We analyze the impact of noise in IMU data on association performance. Zero-mean Gaussian noise is injected into each dimension of IMU data. Then the original IMU vectors are replaced with IMU vectors $\widehat{IMU}$ with noise as both input and output of our model: $\hat{T}^t_{m_{d_i}} = T^t_{m_{d_i}} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2_{d_i})$, where $T^t_{m_{d_i}}$ is the i^{th} dimension vector. $\mu = 0$ and σ_i is the standard deviation of $T^t_{m_{d_i}}$. σ_{d_i} is specific to each dimension due to different range in each sensor (for example, accelerometer vs gyroscope vs magnetometer).

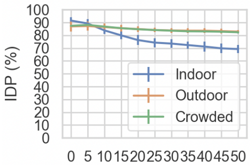


Fig. 9. Effect of IMU noise on association accuracy. We observe that even for high IMU noise, *ViTag* offers reasonable association accuracy.

Noise is injected in different levels from 0% to 50% of the measurement range for each dimension of each IMU sensor. For one test sequence, ten experiments were repeated for each noise level in each dataset. The results are shown in Figure 9 on test sequences with relatively high and low IDPs in *Indoor* (Sequence #3 and #6) and *Outdoor* (Sequence #4 and #9) datasets (0-indexed), respectively. Each vertical line in one dataset denotes a total of 20 experiments (2 sets of 10 repeated experiments) on the aforementioned two test sequences. Apart from noise injection, we follow the same settings in VI-A. Overall, results show that our system is robust to IMU noise.

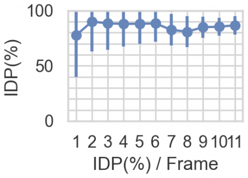


Fig. 10. Impact of the number of pedestrians per frame on association accuracy. We see that IDP is not affected even when multiple persons are detected in one frame.

3) Effect of Number of Detected Person: To see how the number of detected pedestrians in camera view affects association, statistics of per-frame IDP is depicted in Figure 10. Overall, IDP jitters between 80% and 90%, demonstrating that *ViTag*'s performance is robust to the number of detected persons per frame. Note that some people walk out of the camera view in the *Indoor* dataset, which makes the association more challenging (for eg. when number of phones ≥ 1), resulting in higher variance when the number of detections is only 1 per frame.

VII. Related Work

Vision-based Detection and Tracking. Deep neural networks have been employed to detect persons in the camera view and generate trajectories across frames, represented using sequences of bounding box coordinates [4], [35]. These trajectories are more accurate than those using hand-crafted features [22]. We leverage vision trajectories (referred to as tracklets), generated by state-of-the-art vision trackers [4], in the proposed association algorithm.

Inertial Tracking. Inertial sensors have been used in handheld devices. Silva et al. [29] developed and trained an LSTM-based network to reconstruct moving trajectories. IMUs are also capable of profiling user motion and can be used for identification [38]. We exploit motion information obtained from IMU sensors for cross-modal association.

Wireless Ranging. Wireless data has wide application usage, including localization with visual sensors [3], bounding box esimation [20], WiFi ranging measurements [12], etc. Fine Timing Measurement (FTM) protocol (802.11REVmc), introduced in IEEE 802.11-2016 Standard [1], aims to perform wireless ranging with the round trip time (RTT) between an access point (AP) and a WiFi station (STA).

Method	No Pose Used	# IMU Devices / Person	Matching Duration (sec)	# Det / Frame	Avg. Acc (%)
PHADE [7]	-	$\geq$ 1	18	2-10	92.0
ZeroNet [18]	-	1	-	1	82.4
IDIoT [28]	-	13	20	1	92.2
Z-Shot [32]	-	2-4	5.12	1	$<$ 75.0
C. Loc. [13]	✓	1	Cumulative	2-12	82.0
Vi-Fi [17]	✓	$\leq$ 1	1-3	2-11	82.93
ViTag (Ours)	✓	$\leq$ **1**	**1-3**	**2-11**	**88.39**

TABLE II
Summary of Comparison with Representative Works. Note: Det and Acc mean Detected Person and Accuracy, respectively.

Multimodal Association. There has been a large body of research on matching identity across different modalities [2], [9]. The closest work to *ViTag* is Vi-Fi [17] — a deep learning based method that applies affinity losses to learn identity assignment on camera, IMU and FTM readings. Similar to Sun et al. [31] using triplet loss, Masullo et al [23], [24] associate silhouette images and accelerations by deep learning features. A joint representation [25] can be learned to fuse different modalities in an encoder-decoder architecture. In the work by Akbari et al. [2], correlation is done in a handcrafted manner by mapping the acceleration in RGB image plane to physical acceleration, but it fails in limited-varying depth changes, multiple persons, varying brightness or line-of-sight motion direction. Research on vision and IMU association also include PHADE [7], IDIoT [28] [11], [37] etc.

Due to the differences in dataset, methodology, and experimental setup, it is unlikely to draw a direct and fair comparison between the performance of previous work and *ViTag*. We therefore clarify and summarize the distinctions in Table II based on several parameters. Our system *ViTag* performs in much more challenging real world scenarios requiring fewer devices per person, while relying on shorter sequence of measurements, compared to past work.

VIII. Conclusion

In this work, we explored the challenges and solutions associated with cross-modal association. We designed *ViTag* to associate visually detected persons from a camera stream with corresponding smartphone IDs. We proposed *X-Translator* which is a multimodal LSTM-based autoencoder that learns a joint representation between the modalities during training, and translating data from camera modality (vision tracklets) into phone domain (IMU readings and FTM). This enabled us to reconstruct phone data tracklets and match them with the observed phone tracklets in real time to associate identities. Our system achieves an average Identity Precision (IDP) of 88.39% (90.21%, 87.85%, and 87.11% in *Indoor*, *Outdoor*, and *Crowded* datasets, respectively) in an online manner, outperforming the state-of-the-art approach Vi-Fi (IDP = 82.93%) in diverse real-world environments.

Future directions include generalizing *ViTag* to different camera views, leveraging cross-modal attention mechanisms to learn a better joint representation to improve association performance.

IX. Acknowledgement

This research has been supported by the National Science Foundation (NSF) under Grant Nos. CNS-2055520, CNS-1901355, CNS-1901133.

We thank Rashed Rahman,Shardul Avinash, Abbaas Alif, Bhagirath Tallapragada and Kausik Amancherla for their help with data labeling.

References

[1] ”IEEE Standard for Information technology–Telecommunications and information exchange between systems Local and metropolitan area networks–Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications”. *”IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)”*, pages 1–3534, Dec 2016.

[2] A. Akbari, P. Liu, B. J. Mortazavi, and R. Jafari. Tagging wearable accelerometers in camera frames through information translation between vision sensors and accelerometers. In *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*, pages 174–184, 2019.

[3] A. Alahi, A. Haque, and L. Fei-Fei. Rgb-w: When vision meets wireless. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3289–3297, 2015.

[4] P. Bergmann, T. Meinhardt, and L. Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019.

[5] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

[6] A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946.

[7] S. Cao and H. Wang. Enabling public cameras to talk to the public. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–20, 2018.

[8] G. B. Coleman and H. C. Andrews. Image segmentation by clustering. *Proceedings of the IEEE*, 67(5):773–785, 1979.

[9] S. Fang, T. Islam, S. Munir, and S. Nirjon. Eyefi: Fast human identification through vision and wifi-based trajectory matching. In *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 59–68. IEEE, 2020.

[10] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[11] R. Henschel, T. Von Marcard, and B. Rosenhahn. Accurate long-term multiple people tracking using video and body-worn imus. *IEEE Transactions on Image Processing*, 29:8476–8489, 2020.

[12] M. Ibrahim, H. Liu, M. Jawahar, V. Nguyen, M. Gruteser, R. Howard, B. Yu, and F. Bai. Verification: Accuracy evaluation of wifi fine time measurements on an open platform. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, pages 417–427. ACM, 2018.

[13] D. Jung, T. Teixeira, and A. Savvides. Towards cooperative localization of wearable sensors using accelerometers and cameras. In *2010 Proceedings IEEE INFOCOM*, pages 1–9. IEEE, 2010.

[14] W. Krzanowski. *Principles of multivariate analysis*, volume 23. OUP Oxford, 2000.

[15] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[16] H. Kwon, C. Tong, H. Haresamudram, Y. Gao, G. D. Abowd, N. D. Lane, and T. Plötz. Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(3), Sept. 2020.

[17] H. Liu, A. Alali, M. Ibrahim, B. B. Cao, N. Meegan, H. Li, M. Gruteser, S. Jain, K. Dana, A. Ashok, et al. Vi-fi: Associating moving subjects across vision and wireless sensors.

[18] Y. Liu, S. Zhang, and M. Gowda. When video meets inertial sensors: Zero-shot domain adaptation for finger motion analytics with inertial sensors. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, pages 182–194, 2021.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[20] C. X. Lu, M. R. U. Saputra, P. Zhao, Y. Almalioglu, P. P. de Gusmao, C. Chen, K. Sun, N. Trigoni, and A. Markham. milliego: single-chip mmwave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 109–122, 2020.

[21] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi. Mobile sensor data anonymization. In *Proceedings of the international conference on internet of things design and implementation*, pages 49–58, 2019.

[22] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei. Deep learning for visual tracking: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[23] A. Masullo, T. Burghardt, D. Damen, T. Perrett, and M. Mirmehdi. Who goes there? exploiting silhouettes and wearable signals for subject identification in multi-person environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[24] A. Masullo, T. Burghardt, D. Damen, T. Perrett, and M. Mirmehdi. Person re-id by fusion of video silhouettes and wearable signals for home monitoring applications. *Sensors*, 20(9):2576, 2020.

[25] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. Multimodal deep learning. In *ICML*, 2011.

[26] V. F. Rey, P. Hevesi, O. Kovalenko, and P. Lukowicz. Let there be imu data: Generating training data for wearable, motion sensor based activity recognition from monocular rgb videos. UbiComp/ISWC ’19 Adjunct, page 699–708, New York, NY, USA, 2019. Association for Computing Machinery.

[27] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016.

[28] C. Ruiz, S. Pan, A. Bannis, M.-P. Chang, H. Y. Noh, and P. Zhang. Idiot: Towards ubiquitous identification of iot devices through visual and inertial orientation matching during human activity. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 40–52. IEEE, 2020.

[29] J. P. Silva do Monte Lima, H. Uchiyama, and R.-i. Taniguchi. End-to-end learning framework for imu-based 6-dof odometry. *Sensors*, 19(17):3777, 2019.

[30] A. Stisen, H. Blunck, S. Bhattacharya, T. S. Prentow, M. B. Kjærgaard, A. Dey, T. Sonne, and M. M. Jensen. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*, pages 127–140, 2015.

[31] X. Sun, X. Weng, and K. Kitani. When we first met: Visual-inertial person localization for co-robot rendezvous. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10408–10415. IEEE, 2020.

[32] C. Tong, J. Ge, and N. D. Lane. Zero-shot learning for imu-based activity recognition using video embeddings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(4):1–23, 2021.

[33] B. Wang, X. Liu, B. Yu, R. Jia, and X. Gan. Pedestrian dead reckoning based on motion mode recognition using a smartphone. *Sensors*, 18(6):1811, 2018.

[34] H. Wang, X. Bao, R. Roy Choudhury, and S. Nelakuditi. Visually fingerprinting humans without face recognition. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pages 345–358, 2015.

[35] Y. Xiang, A. Alahi, and S. Savarese. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE international conference on computer vision*, pages 4705–4713, 2015.

[36] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo. Deep multimodal representation learning from temporal data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5447–5455, 2017.

[37] J. Zeng, P. Wang, Q. Zhao, J. Pang, J. Tao, and X. Guan. Effectively linking persons on cameras and mobile devices on networks. *IEEE Internet Computing*, 23(4):18–26, 2019.

[38] T. Zhang, M. Karg, J. F.-S. Lin, D. Kulic, and G. Venture. Imu based single stride identification of humans. In *2013 IEEE RO-MAN*, pages 220–225, 2013.