TagAttention: Mobile Object Tracing With Zero Appearance Knowledge by Vision-RFID Fusion

Xiaofeng Shi, *Student Member, IEEE*, Haofan Cai, *Student Member, IEEE*, Minmei Wang, *Graduate Student Member, IEEE*, Ge Wang, *Member, IEEE*, Baiwen Huang, Junjie Xie, *Member, IEEE*, and Chen Qian, *Senior Member, IEEE, Member, ACM*

***Abstract*— We propose to study mobile object tracing, which allows a mobile system to report the shape, location, and trajectory of the mobile objects appearing in a video camera and identifies each of them with its cyber-identity (ID), even if the appearances of the objects are not known to the system. Existing tracking methods either cannot match objects with their cyber-IDs or rely on complex vision modules pre-learned from vast and well-annotated datasets including the appearances of the target objects, which may not exist in practice. We design and implement TagAttention, a vision-RFID fusion system that achieves mobile object tracing without the knowledge of the target object appearances and hence can be used in many applications that need to track arbitrary un-registered objects. TagAttention adopts the visual attention mechanism, through which RF signals can direct the visual system to detect and track target objects with unknown appearances. Experiments show TagAttention can actively discover, identify, and track the target objects while matching them with their cyber-IDs by using commercial sensing devices in complex environments with various multipath reflectors. It only requires around one second to detect and localize a new mobile target appearing in the video and keeps tracking it accurately over time.**

***Index Terms*— Radio-frequency identification (RFID), sensing, mobile tracing, perception fusion.**

I. Introduction

AS THE key components of the Internet of Things (IoT), many moving objects (the 'Things') carry their cyber-identities (IDs) such as unique sequence numbers or network addresses. We study the *mobile object tracing* problem, which allows a mobile system to report the shape, location, and trajectory of the mobile objects appearing in a video camera and identifies each of them with its cyber-ID, even if the appearances of the objects are not known to the system. Mobile object tracing is one essential problem of mobile computing with emerging applications such as cashier-free stores (identify and track the customers and the merchandise in their shopping carts), autonomous cars (identify other vehicles and traffic signs), electronic article surveillance (EAS), virtual/augmented reality, TV motion sensing games, and lost child/object searching. In most of these applications, the appearances of the objects (customers, merchandise, vehicles, lost objects) may not be known in advance to the system, or the objects are of a huge variety whose appearances are too many to learn.

Mobile object tracing requires the following specific tasks.

- *Object detection*: detect each mobile object from the video frames and highlight its shape and boundary.
- *Identify matching*: match each mobile object with its assigned cyber-ID.
- *Movement tracking*: obtain the location and moving trajectory of each target object.

These tasks have been individually studied in many areas including computer vision, wireless sensing, and human computer interaction. For example, computer vision may be able to segment a moving object from video frames – most of these methods require the object's appearance is pre-registered and learned. However, computer vision provides no information about the cyber-ID. Wireless sensing methods can tell the cyber-IDs of the objects in an area but their appearances and detailed behaviors are not known. However, combining these two types of methods and achieving fast speed, cost efficiency, and accuracy are still challenging, especially in many applications where the appearances of the moving objects are not known in advance.

Computer vision is a powerful tool for object classification [22], detection [34], segmentation [18], and tracking [44] from images and videos. Most modern computer vision methods can effectively detect and track objects ***only if*** the object's appearance is pre-registered [1], [44]. For example, a comprehensive and annotated data set is usually required to train these learning models. In addition, these vision based methods can only classify the arbitrary objects with their categorical labels, while they cannot process any cyber-ID information and identify objects with similar appearances. On the other hand, tracking approaches based on RFID can only estimate the coarse location of objects as wireless signals are much less robust to the environmental noises (such as device deviations and unanticipated reflectors) [7], [9], [16], [17], [41], [46]. Thus, they fail to precisely localize the targets and report the object appearances (such as shapes and edges).

An intuitive solution is combining computer vision and RFID technologies to simultaneously obtain the location of the target objects from the visual channel and the identities from the wireless channel [11], [24], [25], [31], [45]. However, existing vision-RFID fusion methods cannot achieve mobile

Manuscript received December 24, 2019; revised August 30, 2020; accepted January 9, 2021; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor M. Li. Date of publication January 26, 2021; date of current version April 16, 2021. The work of Xiaofeng Shi, Haofan Cai, Minmei Wang, and Chen Qian was supported in part by the National Science Foundation under Grant 1717948, Grant 1750704, and Grant 1932447. The preliminary version of this article appeared at the IEEE 27th International Conference on Network Protocols (ICNP), 2019. *(Corresponding author: Xiaofeng Shi.)*

Xiaofeng Shi, Haofan Cai, Minmei Wang, Baiwen Huang, Junjie Xie, and Chen Qian are with the Department of Computer Science and Engineering, University California Santa Cruz, Santa Cruz, CA 95064 USA (e-mail: xshi24@ucsc.edu; hcai10@ucsc.edu; mwang107@ucsc.edu; bhuang21@ucsc.edu; jxie29@ucsc.edu; cqian12@ucsc.edu).

Ge Wang is with Xi'an Jiaotong University, Xi'an 710049, China (e-mail: gewang@xjtu.edu.cn).

Digital Object Identifier 10.1109/TNET.2021.3052805

object tracing with ***zero*** human's assistance. They all require to pre-learn the appearances of the objects, either from a vast and well-annotated dataset that describes the target objects or from users' annotation when the targets initially appear in the scene. If the object appearances are unknown, these solutions are *NOT* able to detect and track the objects from the video and match them with their cyber-IDs. In fact, in many applications the system does not know the appearances of the target objects in advance.

In this paper, we argue that the wireless communication between the target and the reader through the RF channel can essentially *assist* the visual channel to actively find the target mobile object *without* knowing the objects' appearance. We consider the raw visual sensing information (such as video frames obtained from cameras) as the bottom-level information and the abstraction of the objects (such as their cyber-IDs and coarse motion trajectories which can be obtained from the RF channel) as the top-level information. We propose the TagAttention, which adopts the "bottom-up" and "top-down" visual attention model to fuse the visual and wireless sensing channels for mobile object tracing. The "bottom-up" visual attention model predicts the optical flows (patterns of apparent motion of the objects) from the RGB frames and the "top-down" step detects, segments and tracks the visual regions by matching the motion of targets in the video with the tag IDs and wireless channel information. The intention to use attention model in our framework is that physical layer properties of wireless signals, such as signal phases, can "direct" the vision model to focus its attention only to the moving targets. TagAttention could automatically detect, localize, and identify any tagged object in the video when it appears in the camera and then keep tracking it. It only requires around one second to detect and localize a new mobile target appearing in the video and keeps tracking it accurately over time. To our knowledge, **no prior method can achieve this task**.

In summary, the main advantages of TagAttention include: 1) It can actively discover rigid tagged mobile objects and automatically track them without pre-knowledge of the objects' appearance, hence it requires zero human's assistance to label visual data; 2) It is fast and cost-efficient; 3) it does not need manually annotated datasets for training; 4) it uses only commercial off-the-shelf (COTS) devices for sensing and no hardware-level modification is required; 5) it works well in complex and dynamic environments with many multi-path reflectors.

The balance of this paper is summarized as follows. Sec. II presents the related work. Sec. III illustrates the design of TagAttention. In Sec. IV we present the primary evaluation results, and in Sec. V we further analyze the system with empirical studies. The limitations of the system are discussed in Sec. VI. We conclude the paper in Sec. VII.

II. Related Work

A. Localization Based on RFID

Recent RFID research uses physical layer properties of the back-scatter RF signals to localize the RFID tags. The those physical properties typically include received signal strength (RSS) [5], [50], signal phase [26], [27], [36], and angle of arrival (AoA) [43], [51]. However, the accuracy of these methods usually suffers from the multi-path effect caused by destructive reflectors in the environments. In addition, many methods [27], [30], [39] require the RFID tag to be static for a few seconds so that plenty of signal samples are collected for statistical analysis, which makes real-time tracking of the tags in a mobile and dynamic scenario challenging. Meanwhile, the initial measurement bias cause by RFID readers and tags, such as the signal phase bias, usually needs to be carefully measured or canceled before data samples are recorded [27], [43]. Recent studies achieve sub-centimeter localization accuracy by manipulating the radio signals with multiple Universal Software Radio Peripherals (USRPs). For example, TurboTrack [28] estimates the RF signals in much wider bandwidth to reduce the impact of environmental reflectors. However, these methods can hardly be compliable with commercial off-the-shelf (COTS) RFID readers.

In TagAttention, since we can adopt vision as an additional channel which provides plenty of spatial information of the target we want to trace (although with vision alone we do not know exactly the tracing target), the system does not demand a precise localization performance using the RF signals. Hence, the system does not rely on the self-defined radio signals and can be easily implemented with most commercial RF readers and 3D cameras.

B. Visual Tracking Systems

Object tracking in computer vision research is usually defined as predicting bounding boxes for certain objects in every video frame. One category of the solutions uses correlation filters, such as MOSSE [2] filter. More recently, the target patch searching can be accomplished in an end-to-end manner by deep neural networks [23], [48]. Another type of methods utilize the motion information in spatio-temporal context [44] or optical flows [6] of the video, which can also be learned through DNNs. The third type adopts the tracking-by-detection strategy to track specific objects, such as human bodies [14], [15].

However, all the above methods require either a large well-annotated dataset to train their models, or users' initial annotation to tell the model what to track, or both. Actively finding and identifying the targets that are not registered or learned by the models remains unsolved.

C. Vision-RFID Fusion

In recent years, attempts have been made to fuse vision and RF signals so that the systems can both track and identify mobile targets by matching the information from both channels [11], [12], [24], [25], [31], [45]. Mandeljc *et al*. [31] propose to detect and track anonymous humans from videos with Probabilistic Occupancy Map (POM) algorithm, and then identify the individuals by matching the IDs in RF-channel to the detected human instances based on the location information. ID-Match [25] is a novel vision-RFID fusion system for human identification from a group through an RGB-D camera and an RFID sensor. However, both of the above-mentioned methods rely on the human detection or human pose estimation module accomplished by specifically-trained computer vision models. Therefore they cannot be used to identify objects other than humans.

Beyond tracking and identification of humans, TagVision [11], [12] fuses signals of RFID tags on objects and 2D surveillance video by calculating probabilistic matching scores of the signal phases and object motions. However, the vision

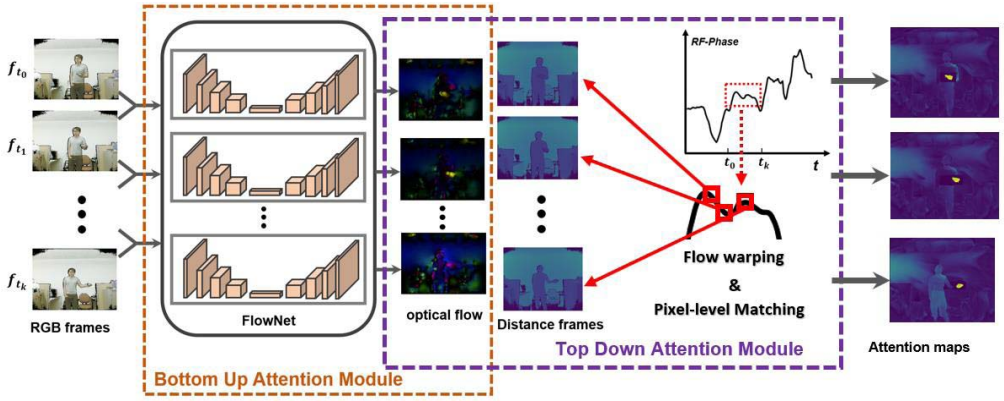


Fig. 1. Overview of TagAttention. The system is mainly comprised of the bottom-up and top-down attention modules.

model is hard to be applied in complex 3D scenarios: it can only track objects on a static 2D plane by which the camera model is calibrated. A recent work proposes IDCam [24], which fuses RFID and 3D camera to trace a tagged item that is held by a user's hand. The system requires a precise detection of the user's gestures, which is accomplished by a carefully tuned visual detection and tracking module. In addition, TaggedAR [45] is proposed to detect and identify stationary objects by rotating the sensors and pairing RF-signals with the depth of the target objects. However, the system discards the informative object descriptions from the visual intensity channels and simply segments objects from the background based on depth histogram, which significantly reduces the robustness of the system in complicated scenarios.

Existing fusion solutions cannot achieve tracing arbitrary mobile objects in 3D space. They either only trace particular targets (such as a human body) with sophisticated models or trace objects on a calibrated 2D plane. They cannot identify and track objects with unknown arbitrary appearances in complex 3D environments, which is our design objective of this work.

III. Design of TagAttention

A. Overview

In TagAttention, we use a commercial RFID reader carrying one antenna and an RGB-D camera on top of the antenna to capture the sensing data. In addition, each tracing target carries an RFID Tag that can be read by the RFID reader through the antenna. Fig. 1 shows an overview of our attention-based fusion system. The inputs of our fusion model are the RGB intensity and distance maps (each pixel of the distance map represents the distance from the 3D voxel to the sensor origin) captured by the RGB-D camera, and the RFID EPCs (denoting the cyber-IDs of the objects) and their corresponding phase signals obtained by the RF reader.

We consider the raw video inputs as the bottom-level information and the abstraction of the objects (such as their cyber-IDs and motion trajectories) as the top-level information. Given two consecutive RGB frames, the *bottom-up* visual attention mechanism estimates the pixel-level optical flow to measure the motions of pixels from the visual frames. Since the produced optical flow can highlight moving pixels from raw video, it works as a bottom-up visual attention mechanism [8], where the system naturally notice the salient visual components of potential importance from visual inputs.

Meanwhile, the *top-down* visual attention module in TagAttention functions as a detector of the targets given the RF signals that match the visual targets. In the top-down attention module, we obtain the consecutive distance by unwrapping the phases of RFID tags, and map it with the per-frame optical flows. By combining the *bottom-up* and *top-down* modules together, we can obtain an attention map for each timestamp, which represents the pixel-level consistency between the motion trajectories in the video and the distance changing of the RFID tag. The attention map is a 2D matrix with the same size as the video frame resolution, in which each element represents the magnitude of attention (measured by the probabilistic matching score of the two sensing channel in our design) on the corresponding image pixel.

Finally, a tracker is designed to actively discover the target objects and output their corresponding shape and location (represented by a pixel-wise mask for the object, we use 'mask' in the following) from the video based on the per-frame attention maps.

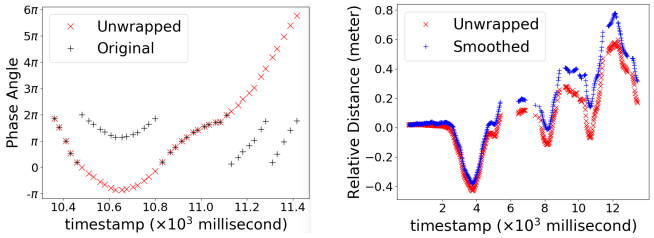
Compared to the existing fusion methods, TagAttention can actively highlight ubiquitous target objects in a video *without* any pre-knowledge of the object's appearance. Thus, this tracing model can be applied on a much wider variety of visually-complex scenarios in which target objects are not visually pre-registered.

B. RF Signal Preprocessing

In TagAtthention, the RFID tags are matched to the objects in the video through the correlation of the motion trajectories of the objects. The distance L from the reader antenna to the tag can be calculated as follows:

$$L = \frac{\phi_L \cdot c}{4\pi f}, \tag{1}$$

where ϕ_L represents the corresponding phase change over the signal travel distance, c is the speed of light and f is the signal frequency (equals to 920MHz for our reader). Note that with the current COTS devices, we can not calculate the exact distance of the tag. There are two reasons. One is that in addition to the phase ϕ_L over distance, both the reader and tag's circuits will introduce some additional phase rotations to the received phase ϕ, i.e., $\phi = (\phi_L + \phi_R + \phi_T) \mod 2\pi$, where ϕ_R and ϕ_T are the additional phases of the reader and tag respectively [13], [19]. Another reason is that our commercial RFID reader (ImpinJ R420) also introduces π radians of ambiguity. In other words, the reported phase can either be the true phase or the true phase plus π radians [19]. Hence for our reader, $\phi_L = n\pi + \phi - (\phi_R + \phi_T)$, where n is a non-negative integer. Since ϕ_R and ϕ_T are constant over



(A) De-periodicity (B) Smoothing

Fig. 2. RF phase signal preprocessing and the relative distance trajectory.

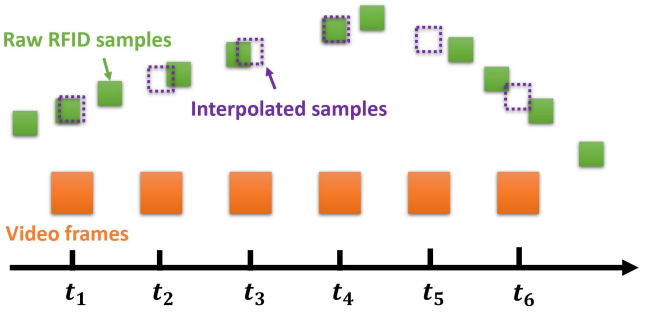


Fig. 3. Channel synchronization and sensing data Sampling.

the whole reading period, to estimate the motion of the tag over time, we only consider the relative distance changes of the tag, i.e.,

$$\Delta L = L - L_0 = \frac{(\Delta n\pi + \Delta\phi)\cdot c}{4\pi f}, \tag{2}$$

where L_0 is a reference distance which can be set as the first calculation. And $\Delta n = n - n_0$ and $\Delta\phi = \phi - \phi_0$. After this step, we can obtain a relative moving distance of the tag, ΔL, which only related to the changing positions.

To extract the motion trajectory of the objects, we conduct two signal processing progress, namely phase de-periodicity [4] and motion smoothing. As illustrated as the black plus sign in Fig. 2 (A), the received phases are wrapped over cycles and fall into the range of 0 to 2π. This characteristic of the phase values makes the motion estimation discontinuous. Hence we first unwrap the received phase values and retrieve the consecutive motion profile. In our design, we adopt two thresholds, $th_1 = 0.5\pi$ and $th_2 = 1.5\pi$, to detect the π and 2π hops. Specifically, let $\Delta\phi_{t_1,t_2} = |\phi_{t_2} - \phi_{t_1}|$ represent the difference between two adjacent phases ϕ_{t_1} and ϕ_{t_2}. The latter phase value ϕ_{t_2} will be added or subtracted by π if $th_1 < \Delta\phi_{t_1,t_2} \leqslant th_2$, and by 2π if $\Delta\phi_{t_1,t_2} > th_2$. The performance can be found in Fig. 2 (A).

We also consider the motion smoothing to get rid of the environment and device noises. Since the received phases can be easily impacted by outside environments and equipments, it is hard to tell whether a hop between adjacent received phases is caused by the π or 2π phase wrapping, or by a sudden movement of the object, or by insufficient reading. Hence, we further smooth the phase based on the estimated acceleration of the moving object. The main idea is based on an observation that the rapid and sudden change of velocity, which requires a huge force acting on the object, is unlikely to happen in most real applications. Thus, we calculate the average velocities and accelerations of the object within the reading time slots after de-periodicity. If the acceleration of the object in a certain time slot is higher than a threshold, i.e. the gravity acceleration $g \approx 9.8m/s^2$, we consider the high acceleration is caused by the inappropriate de-periodicity or other environmental noises. To smooth the motion of the objects in such case, we keep the average velocity $\bar{v}_{t_0,t_1}$ in previous time slot constant for the next time slot and approximate the gain of distance at t_2 by $(t_2 - t_1)\bar{v}_{t_0,t_1}$. A smoothing result is shown in Fig. 2 (B).

C. Channel Synchronization

The fusion of the RFID and Vision channels requires the synchronization of two-channel data samples. When collecting the data from the two channels, we use the operating system's clock to generate a timestamp signature for each of the RFID phase sample and video frame.

When there are multiple targets and interference RFID tags in the same scene, the phase sampling rate of each RFID tag becomes rather non-uniform. First, due to the uncertainty in slotted ALOHA protocol, we cannot predict which tag will respond and occupy the next slot. Second, since most of our experiments are conducted in a noisy environment and the target object is placed at a relatively long distance (2 to 5 meters) from the antenna, many packets carrying the target tag information may get lost during transmitting. Therefore, to synchronize the two channels, we use the timestamps of Kinect data as the timestamps for channel fusion. Then as shown in Fig. 3, the tag distance trajectory obtained from the RFID channel (as shown in Fig. 2) is interpolated and resampled to match each Kinect frame. Specifically, we use polynomial interpolation in our implementation.

D. Bottom-up Attention Module

In TagAttention, the bottom-up attention module captures the salient visual features through the optical flow, i.e. the motion of pixels in two consecutive video frames at t and $t + \Delta t$. The optical flow will be used to warp[1] the video distance maps and propagate the predicted attention maps over frames.

In our framework, we learn the optical flow through an end-to-end deep neural network, which has been proved to be both more effective and efficient [10], [32] than traditional methods. Specifically, we adopt the FlowNet [10] as the backbone neural network architecture and the training strategy presented by [32] to train the neural network in an unsupervised manner.

By feeding the consecutive video frame pairs F_{t_1}, F_{t_2} into the FlowNet, the model predicts the optical flow map $f_{t_1\to t_2} = \{(\Delta x, \Delta y)\}_{(x,y)}$. The estimated optical flow naturally highlights the pixels on moving objects from the image frames, which works similarly as a visual bottom-up attention mechanism to notice the mobile objects. In addition, the optical flow will be further used to warp the distance maps and propagate the predicted attention maps over frame timestamps. Note that the FlowNet can be replaced with any optical flow model that yields better accuracy.

[1]In this paper, warping stands for forward warping with the optical flow. Namely, we move each pixel of the current frame in the image plane according to the pixel velocity, such that we can reconstruct a "virtual" frame for the next timestamp.

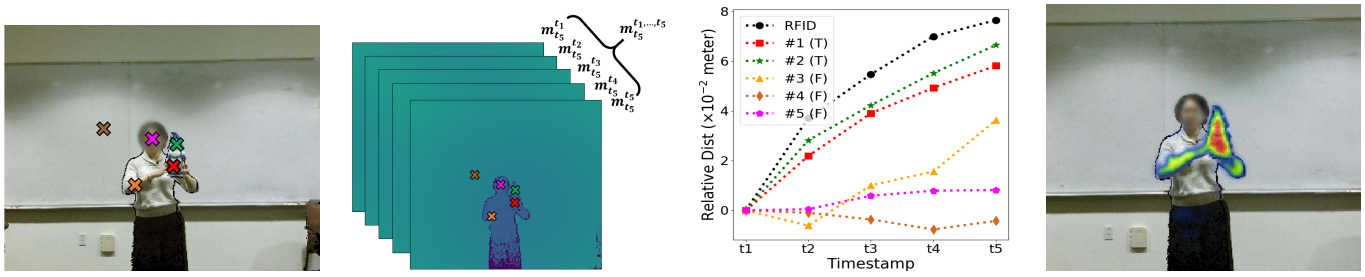


Fig. 4. (A): Samples of the anchors in an example video frame. (B): The corresponding motion map of the frame in (A) (window size = 5). (C): The motion trajectories of the anchor points: Pixel #1 and #2 are the target anchor pixels, while the rest are random anchor pixels. (D): The generated attention heat map.

E. Top-Down Attention Module

1) Motion Estimation: In the top-down attention module, TagAttention finds and highlights the target objects' pixels by matching the motion of each pixel in the visual system with the distance changes measured by the RF phase and calculating their correlation probabilistic scores. To estimate the pixel-level motion (the moving trace of each pixel in Kinect frames), we warp the distance maps D_{map} with the optical flows frame by frame and obtain the motion maps M_{map}. In M_{map}, each pixel denotes the distance trajectory (represented by a vector) of the invariant real-world voxel in 3D space. Specifically, let $d_0, d_1, \ldots, d_t \in D_{map}$ represent distance maps from the first frame F_0 to the current frame F_t, and $f_{0\to1}, f_{1\to2}, \ldots, f_{t-1\to t} \in Flow_{map}$ represent the optical flows of the RGB video. We warp D_{map} with $Flow_{map}$ to estimate the motion maps M_{map}. Let $m_{t_j}^{t_i,\ldots,t_j}$ be a instance of M_{map} on frame t_j. The size of $m_{t_j}^{t_i,\ldots,t_j}$ is $H\times W\times(t_j-t_i+1)$, where H and W are the height and width of the video frames, and the third dimension is the time channel from t_i to t_j. Then $m_{t_j}^{t_i,\ldots,t_j}$ can be calculated as Eq. 3:

$$m_{t_i}^{t_i,\ldots,t_j} = ((((d_{t_i} \otimes f_{t_i\to t_i+1}) \oplus d_{t_i+1}) \otimes f_{t_i+1\to t_i+2}) \oplus d_{t_i+2} \cdots \otimes f_{t_j-1\to t_j}) \oplus d_{t_j}, \quad (3)$$

where $\otimes$ represents the warping process with optical flow over all channels of the third dimension of the matrix, and $\oplus$ represents concatenating of two maps along the third channel (i.e. the time channel).

Meanwhile, the RFID reader collects the RF signal for each tag id_k during t_i to t_j, and the signals are converted into relative distance vectors $rd_{id_1}^{t_i,\ldots,t_j}, rd_{id_2}^{t_i,\ldots,t_j}, \ldots, rd_{id_n}^{t_i,\ldots,t_j} \in RD^{t_i,\ldots,t_j}$. We then match the moving pixels with the RF tag by calculating the correlation probabilistic scores between the motion map $m_{t_j}^{t_i,\ldots,t_j}$ and the RF distance vector $rd_{id}^{t_i,\ldots,t_j}$. Fig. 4 presents an example. As shown in Fig. 4, (A) shows an RGB frame at time t_5, and (B) represents the motion map $m_{t_5}^{t_1,\ldots,t_5}$ over five timestamps from t_1 to t_5 ($\approx 150ms$) computed by Eq. 3. In Fig. 4 (A), we arbitrarily sample a few pixels as random anchors and illustrate their motion trajectories in (C). As a comparison, we also label two pixels (denoted by red and green) on the target object as target anchors in $m_{t_j}^{t_i,\ldots,t_j}$ and show their estimated relative distance vectors as well over time in (C).[2] In addition, the motion estimated by RF channel $rd^{t_i,\ldots,t_j}$ is also plotted with the black line in (C). To eliminate the overall bias caused by the π or 2π rotations of RF signal phases, the motion vectors are translated so that the initial relative distance of motion trajectory in the window is 0, namely, for each timestamp t_k within $[t_i, t_j]$, $\hat{m}_{t_j}^{t_k} = m_{t_j}^{t_k} - m_{t_j}^{t_i}$ and $\hat{rd}_{id_n}^{t_k} = rd_{id_n}^{t_k} - rd_{id_n}^{t_i}$. Hence, we obtain the unbiased motion map $\hat{m}_{t_j}^{t_i,\ldots,t_j}$ and RF motion vectors $\hat{RD}^{t_i,\ldots,t_j}$ for comparison and matching (as shown in Fig. 4 (C)). From Fig. 4 (C), we notice the motions of the two anchor pixels located at the target object in the motion map match well to the motion of the RFID tag estimated by RF signals, while other random anchor pixels fail to match.

Ideally, the motions of the pixels on a rigid target in the unbiased motion map $\hat{m}_{t_j}^{t_i,\ldots,t_j}$ from the visual channel should perfectly match with the unbiased motion vector of the corresponding RFID tag, since they all measure the relative distance from the anchor point of the object to the sensors within timestamp t_i to t_j in the physical 3D space. However, both measurements could be inaccurate, causing the possible misalignment of the two traces. For example, in the visual channel, error exists when warping the distance map as the optical flow may not be perfect; while in the RF channel, the error can be caused by multi-path, random Gaussian noise, low sampling rate and inappropriate De-periodicity. Nevertheless, the tendency of the motions in two channels can match in a long term, as all these noisy factors only cause random and temporary impact on the signals. Hence, we introduce an attention mechanism Att_{RF}, which is robust to the temporary and random noise, to measure the correlation of the motions in different channels.

2) Attention Mechanism: The proposed attention mechanism Att_{RF} is comprised of two attention components: 1) Att_{rbf}, which uses an radial basis function (RBF) kernel to measure the similarity of the motion vectors in Euclidean space; 2) Att_{corr}, which measures the correlation coefficient of the motion vectors. To calculate the attention scores, we first reshape $\hat{m}_{t_j}^{t_i,\ldots,t_j}$ into $\left\{\rho_{(h,w)}^{t_i,\ldots,t_j}\right\}_{H\times W}$, with each element $\rho_{(h,w)}^{t_i,\ldots,t_j}$ representing the motion vector from t_i to t_j of each pixel $p_{(h,w)}$ in the motion map $\hat{m}_{t_j}^{t_i,\ldots,t_j}$. Then the pixel-level attention mechanism can be formulated by Eq. 4 and Eq. 5.

$$Att_{rbf} = exp\left(-\frac{\left\|\rho_{(h,w)}^{t_i,\ldots,t_j} - \hat{rd}_{id_k}^{t_i,\ldots,t_j}\right\|^2}{2\alpha}\right), \quad (4)$$

$$Att_{corr} = Relu\left(\frac{cov(\rho_{(h,w)}^{t_i,\ldots,t_j}, \hat{rd}_{id_k}^{t_i,\ldots,t_j})}{\sigma(\rho_{(h,w)}^{t_i,\ldots,t_j})\sigma(\hat{rd}_{id_k}^{t_i,\ldots,t_j})}\right), \quad (5)$$

[2]Note that the anchors are artificially selected only for the visualization and illustration purpose.

where we use the rectifier activation function $Relu(x) = max(0, x)$ to suppress negative correlations, α is the RBF kernel parameter, $cov(\cdot)$ represents the covariance of the two vectors and $\sigma(\cdot)$ represents the variance of the vector. To combine the two types of attention mechanism together, we used Eq. 6, which calculates the weighted sum of the two attention scores.

$$Att_{RF} = \beta Att_{rbf} + (1 - \beta) Att_{corr}, \beta \in [0, 1] \quad (6)$$

We empirically set $\alpha = 5 \times 10^{-4}$ and $\beta = 0.8$ in our implementation. According to the formulas, Att_{RF} is in the range of $[0, 1]$. Hence we approximately consider Att_{RF} to describe the probability that the pixel (h, w) at timestamp t_j matches with the target object that is labeled by a certain RFID tag. Thus, for each target object, we construct the attention map matrix a_t, which is of the same size as the input image matrix. Each element in a_t represents the attention probabilistic score Att_{RF} of the corresponding pixel. Fig. 4 (D) shows an example of the attention map with a heat map.

F. Attention Propagation

The top-down attention module enables the system to predict an attention probabilistic map for each video frame. However, the prediction can be accurate only when the target objects move during the attention window, since we assume the top-down attention is triggered based on the movement of the targets. When the target object is static, the distance values of the object pixels keep unchanged in the RGB-D camera. However, due to the dynamical factors of the environment (such as the movement of other objects), the phase values of the corresponding tag may still subtly change over time. In such case, the noise of the environment dominates the attention probabilistic scores of the pixels according to Eq. 4 and Eq. 5. In addition, distance measurement or localization of objects through RF signals within a pixel level error bound (about several millimeters) is rather challenging [3], [7], [29], [42], [49], especially when using commercial RFID readers and a single antenna in our system [33], [35], [36], [39], [41], [47]. Therefore, it is nearly impossible to precisely match every pixel with the corresponding RF signals based on the relative motion at a single frame. Fortunately, the visual channel provides tremendous semantic information of the target objects and the environments, which enables us to track and segment the target objects cross multiple timestamps based on the correlation of objects' appearances. Though there maybe some mismatches at a few frames, the overall trend of motions of the two channels can finally match with each other in a long term.

Hence, in order to improve the robustness of our tracking system, we propose an Attention Propagation mechanism as illustrated in Fig. 5. The major intuition in the Attention Propagation module is utilizing a history of the continuous frames of the attention maps to learn the shape and position of the target. The historical motion information of the mobile target has been used in many existing RFID tracking systems, for example, in TurboTrack [28], tracking a mobile RFID tag is formulated as a Hidden Markov Model (HMM). However, the HMM method is not a suitable solution in TagAttention for following reasons.

First, the goal of TagAttention is tracing the target "object" instead of the RFID tag (Existing methods consider the tag position as a coordinate in space). Namely, TagAttention

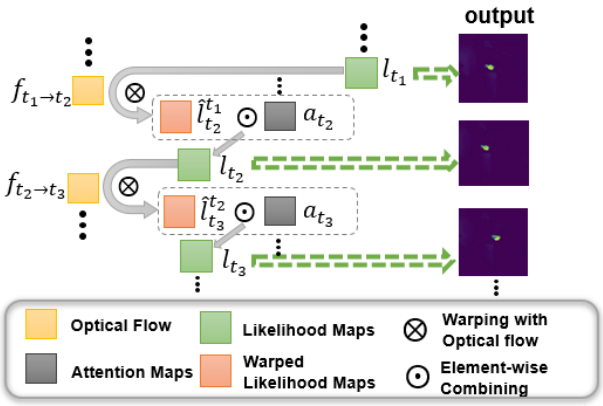


Fig. 5. Mask propagation by warping the probabilistic maps with optical flows over time.

requires not only estimating the motion trajectory of the RFID tag but also detect the shape and position of the associated target by highlighting all target pixels in video. Therefore, it is rather expensive to form an HMM of each candidate pixel in the video. Instead, plenty of historical information about the target positions can be extracted from the optical flows. Compared with HMM, the optical flows learned from a deep neural net can estimate a pixel-level historical trace by considering not only the spatial likelihood of the pixel positions (given the prior positions) but also the similarity of the visual features (such as color intensity and object edges) between the two neighboring frames.

Thus, in the proposed Attention Propagation module, we consider that a model based on optical flow is a more efficient and effective method than the HMM to formulate the historical information and trace the target. Specifically in Attention Propagation (Fig. 5), for each target object instance id_k, we initialize the likelihood map $l_{t_0} = \log a_{t_0}$ ($\log a_{t_0}$ represents the element-wise log operation of the attention map matrix a_{t_0} in our notation) at the first frame F_{t_0}. For each following frame F_{t_i}, we warp the likelihood map l_{t_i} with the optical flow $f_{t_i \to t_{i+1}}$ to reconstruct the warped likelihood map prediction at frame $F_{t_{i+1}}$, which is denoted as $\hat{l}^{t_i}_{t_{i+1}}$. Then the likelihood map $l_{t_{i+1}}$ at frame $F_{t_{i+1}}$ is calculated by Eq. 7,

$$l_{t_{i+1}} = \hat{l}^{t_i}_{t_{i+1}} + \Theta(v_{t_{i+1}} - v_0) \times \log(a_{t_{i+1}}), \quad (7)$$

where $v_{t_{i+1}} = \frac{\left| rd_{id}^{t_{i+1}} - rd_{id}^{t_{i+2-k}} \right|}{t_{i+1} - t_{i+2-k}}$ denotes the absolute velocity of the motion of the target measured by the RF signal within the time window in which $a_{t_{i+1}}$ is computed, k is the window size (count of the timestamps in the window), $\Theta(x) = 1$ if $x > 0$ otherwise $\Theta(x) = 0$, and $v_0 > 0$ represents a velocity threshold.

In our implementation, we set $v_0 = 0.1\ m/s$, meaning the system is only triggered by the mobile targets that move at a temporary absolute velocity higher than $0.1\ m/s$. An empirical analysis about the parameter v_0 will be discussed in Sec. V-B.

G. Tracking by Attention

In the previous attention modules, only the pixels of the target object in video frames would have consistently high attention probabilistic score over different timestamps, thus yielding high likelihood value in the current likelihood

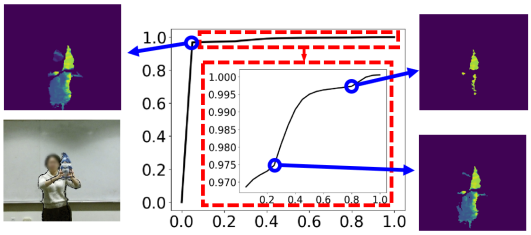


Fig. 6. CDF of the normalized probabilistic values in an example probabilistic map p_{t_i}. Blue circles represent the Conner points in the CDF plot.

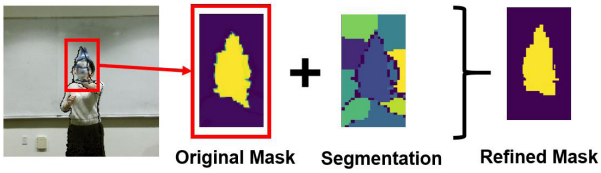


Fig. 7. Mask refinement.

map l_{t_i}. Therefore, we can simply use a threshold to cut off the likelihood and segment the target in current frame F_{t_i}. However, according to Eq. 7, the likelihood value of each pixel keeps decreasing over time as more frames are processed, which makes it infeasible to set a fixed cutting-off threshold. Therefore, we design an automatic thresholding method to segment the target from the video frames based on the likelihood map.

Specifically, we first convert the likelihood map l_{t_i} to the normalized probabilistic map p_{t_i} by calculating $p_{t_i}(h,w) = e^{l_{t_i}(h,w)}$ in element-wise of the 2D matrix l_{t_i}. Then we normalize p_{t_i} crossing all pixels using min-max normalization. By observing the value distribution of the pixels in the probabilistic map p_{t_i}, we can easily find that the probabilistic values are highly hierarchical: the background pixels, which usually comprise the major regions of the frame image, have significantly smaller probabilistic values (close to 0) than the target objects; the"soft" body components that temporarily move in consistency with the target rigid body would have relatively smaller probabilistic values, and the values of these body pixels keep decreasing when the motion consistency no longer holds; while the target object would have consistent highest values. Fig. 6 shows an example of the cumulative distribution function (CDF) of the pixel values in p_{t_i}. Based on this observation, we can use multiple ways to segment the frames according to the normalized probabilistic map, such as value clustering or simply cutting off the CDF of the value distribution at the "corners" (showing as a sudden change of the gradient) on the CDF plot (as labeled in Fig. 6). In our implementation, we choose the last corner point in the CDF to cut-off the image to extract the target mask.

Another issue of the tracking system is that the errors in the predicted optical flow accumulate over the warping steps, resulting in the possible misdetection of the target after a few iterations of attention propagation. To solve this problem, we refine the shapes of the target masks according to the 3D segmentation of scene based on K-means clustering [20], [38]. Fig. 7 illustrates an example of the segmentation and refinement. Then the refined likelihood maps are used in Eq. 7 for attention propagation.

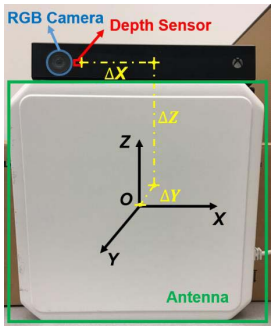


Fig. 8. Deployment of sensors.

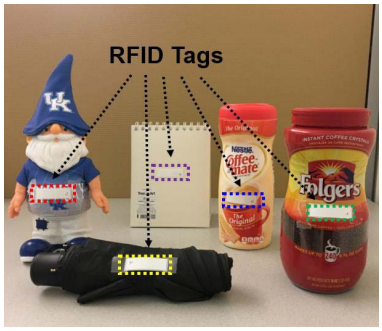


Fig. 9. Examples of target objects.

IV. Evaluation

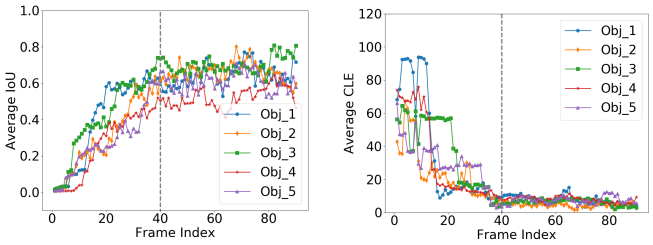
A. Implementation

In our experiments, we utilize a similar sensor setting as [25] to obtain the visual frames and RFID signals. As shown in Fig 8, a Kinect v2 camera is deployed on the top of an RFID antenna. The antenna is connected to a commercial RFID reader ImpinJ R420. We choose the center of the antenna as the origin O of 3D localization reference system and measure the coordination $(\Delta X, \Delta Y, \Delta Z)$ of the depth sensor on the Kinect. Thus, the XYZ 3D point cloud in Kinect reference system could be translated by $(\Delta X, \Delta Y, \Delta Z)$ to obtain the coordination of pixels in the RF reference system.

In our implementation, the FlowNet [10] module for optical flow estimation is implemented with Tensorflow, and we used the loss functions and parameter settings suggested by [32] for training. The neural network is first pre-trained on the synthetic dataset FlyingChairs [10] without using the ground truth data, then fine-tuned on Kinect video frames collected arbitrarily in dynamic environments. The Top-down attention module is also implemented jointly with FlowNet in Tensorflow, but no training is required for this part. The whole system is tested with one Titan X GPU and 8 vCPUs @ 2.6 GHz. Without any decent optimization in the implementation, the average overall processing time for each video frame is around $95ms$, which demonstrates the potential of the proposed method to be applied to online tracking systems.

B. Experiment Setup

To evaluate the performance of the tracing system, we ask 2 volunteers to move everyday objects continuously with arbitrary traces in front of the sensors. Examples of the objects that we tested are shown in Fig 9. The objects tested are of different shapes, sizes, materials and textures. We stick an



(A) IoU of single target (B) CLE of single target

Fig. 10. The petracing. Panel (A) and (B): Average IoU (A) and CLE (B) of each tested target object.

RFID tag on each of the objects. When collecting sensing data, the Kinect records the RGB image frames and 3D coordination of the pixels. Meanwhile, the RFID reader records the tag EPCs (considered as the cyber IDs of the targets) and phase information.

Tracing cases: We consider two tracing cases in our evaluation, namely *single moving target tracing* and *multiple moving targets tracing*. In the single moving target tracing case, we conduct the experiments in two totally different environments. One is in a relatively static meeting room with several furniture (e.g., tables and chairs) in it. In this environment, we test tracing of 5 different objects and repeat for 4 times for each object. Besides, to investigate the impact of noise factors such as multipath effects of the RF signals, we also evaluate our system in a noisy and crowded office room, which has narrow open space, multipath reflectors (tables, chairs, cubicle walls), metal and electronic furniture (cabinet, servers, workstations), various wireless signals (WiFi, LTE), and magnetic fields (whiteboard) in it. We also ask another volunteer to keep walking around to make some dynamic noises. The experiment in such scenario is repeated for 5 times.

We also evaluate the system for tracking multiple moving targets and assign the correct ID to each of them in a noisy environment (the office room scenario). Some of the tested targets are of the similar appearance. Thus, a pure vision-based detection system cannot distinguish them.

C. Evaluation Metrics

We use the Intersection over Union (IoU) and Center Location Error (CLE) to evaluate the tracing performance. IoU is calculated as Eq. 8:

$$IoU = \frac{S(B_t \cap B_p)}{S(B_t \cup B_p)}, \tag{8}$$

where $B_t \cap B_p$ and $B_t \cup B_p$ represent the intersection and union of the ground truth bounding box B_t and predicted bounding box B_p of the target object in video frames respectively, and $S(X)$ represents the area of the region X. CLE measures the Euclidean distance (in number of image pixels) between the centers of the ground-truth bounding box and predicted bounding box in pixels, compared with to the overall input/output frame resolution 512×424.

D. Single Object Tracing

1) Tracing in Static Environment: Fig 10 shows the performance of tracing single target in static scenarios. In Fig 10, plot (A) and (B) show the average IoU and CLE metrics of the five different target objects respectively, where the X axis represents the timestamps of the 90 video frames, and the Y axis represents the average IoU or CLE value.

The evaluation results in Fig 10 illustrate the process in which TagAttention gradually and actively discover the targets and keep tracking them over time. We find TagAttention achieves low IoU scores and high center location errors in the first 20 video frames (at the very beginning frames, the IoUs are always close to 0), showing initially TagAttention cannot track anything as it knows little information about what to trace. This property contrasts to the existing tracking systems, in which they find the targets' location well at the initial stage by human's assistance or an object detection module that is well-trained on large datasets to learn the target. However, we notice the IoU score keeps increasing and the error keeps decreasing until around the $40th$ frame, showing TagAttention can gradually find the location of the targets based on the consistency of the target motion trajectories observed from both sensing channels. Moreover, after around the $40th$ frame, TagAttention becomes confident of the objects' location and mask. Then it keeps tracking the objects for the following frames, yielding high IoUs, low CLEs.

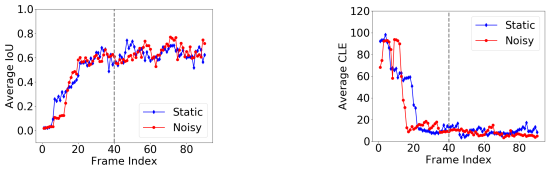
In Fig 11, we present examples of the attention heatmaps a_t learned by the top-down attention module. Warmer color in the figure represents higher attention probabilistic score. Image regions that are not masked by the heat map have 0 attention score. The number at left-up corner of each image shows the frame index. Due to the error factors including the channel noise, inaccurate motion warping and multipath of the signals, or even the negligible motion velocity of the target at certain frames, we can find in several frames the top-down attention module cannot always only focus attention on the target. However, the target can always receive continuous and stable attention from the tracer for most of the frames, enabling TagAtthention to trace the target in a long term.

2) Tracing in Dynamic and Narrow Environments: To evaluate the impact of environmental noises, such as multipath effects, to our tracing system, we conduct the tracing experiments in a dynamic and crowded office room. Fig 12 shows the performance in comparison with the tracing results of the same target in the previous static environment.

Fig 12 (A) and (B) shows the average IoU and CLE results respectively. From the results, we notice the tracing performances in two different scenarios are equivalent, which shows the system is robust to multipath of the signals. In fact, since TagAttention only estimates the coarse motion of the targets rather than accurate localization using the RF signals, the system does not suffer as much from inaccurate phase measurement. In addition, the smoothing methods introduced in Section 3.2 to preprocess the RF signals and the mask refinement strategies introduced in Section 3.6 also help to minimize the impact of signal noise in real-world scenarios.

To better illustrate the actual tracing quality and investigate where the errors come from, we show some selected tracing results of the single object scenarios in Fig 13. Specifically, the first row in Fig 13 shows the meeting room scenario, the second row shows the office scenario, and the third row shows how the system reacts with errors that occur at certain frames. In Fig 13, the number at the left-up corner of each image indicates the frame index in the tracing scenarios. The IoU and CLE of the tracing performance are also presented below each frame image. From Fig 13, we find most tracing

Fig. 11. Examples of the attention heat maps predicted by TagAttention.



(A) IoU (B) CLE

Fig. 12. Tracing performance of single target in noisy environments.

errors is caused by the ambiguous boundary between the target and its surroundings. Since TagAttention requires no prior knowledge of the appearance of the target, it cannot distinguish the target and its surrounding body parts (i.e. the hand and wrist of the volunteer) that move consistently with the target. In these cases, the system considers the target as well as part of its surroundings as an entire rigid body. Since the bounding box IoU score is sensitive to the redundant areas, especially for small objects, we observe a low IoU score for these predictions, whereas the tracing performance is still acceptable.

From the third row of Fig 13, we also notice that a sudden decrease of tracing performance occurs at the $62nd$ frame after TagAttention has already found an accurate position of the targets. We find this phenomenon happens occasionally during tracking. It is mainly caused by the flow warping error in the tracking module of TagAttention. Usually, in such cases, the optical flow measured by FlowNet is inaccurate at a certain frame. Consequently, when propagating the attention maps, the target image region "leaks the attention values" to some irrelevant image pixels. Then in the mask refinement module, the tracer mistakenly considers these irrelevant pixels are of the same rigid body as the target object because these pixels are also spatially close the target. Hence, it starts tracking more body parts than the target rigid body (for example, the entire human body in frame # 64 in the last row of Fig 13). However, after a few frames, as the irrelevant body parts move inconsistently with the target, the attention values of corresponding pixels decrease quickly. Then the tracer can recapture the accurate position of the target and track only the target part (for example, the $66th$ and $68th$ frame in the last row of Fig 13).

E. Multiple Object Tracing

TagAttention can trace multiple mobile targets simultaneously by their cyber IDs without introducing much extra computation. In fact, the most computationally intensive part in TagAttention is the optical flow module, which estimates the optical flow map through a deep neural network. However, the optical flow of the video can be reused by any top-down attention parts to detect and track different targets. Specifically, when the RFID tags of multiple targets are detected, their EPCs and the corresponding phase signals are recorded and processed independently. After the optical flow and the pixel-wise motion map of the video frames are calculated, TagAttention can use these phases signals to compute the attention values of the pixels and produce their corresponding likelihood maps in parallel.

We evaluate the performance of TagAttention in multiple target tracking scenarios. Fig 14 shows the average IoU and CLE scores of different targets in the two-object tracing scenarios. From Fig 14, we find the performance of TagAttention for each individual target is similar to the single object tracing cases. Specifically, the tracer takes less than 35 frames to discover the accurate location of each individual targets and keep tracking them for the following frames.

In addition, we show some selected tracing frames of two-object and four-object tracing scenarios in Fig 15. At the $5th$ frame, the tracer cannot recognize and detect any targets. After more motion data is collected, TagAttention produces fine-grained bounding box and segmentation mask for each target, and labels the targets by the corresponding tag IDs. Especially in the four-object scenarios, we find the system can distinguish the two cylindrical bottles (ID_2 and ID_3) by their IDs, even though the two bottles are very similar in appearance.

V. System Analysis and Discussion

A. Impact of Moving Speed

To evaluate the impact of the moving speed rate of the target, we design the following human tracing experiment: the volunteer who wears the RFID tag walks toward the sensors in the crowded office room at different levels of speed rates (the average speed rates are about $2m/s$, $1.3m/s$, $0.8m/s$, $0.5m/s$ respectively). Different from previous works [25], [45] that fuse RF signal and vision to track humans, our system does not rely on any human detection module (which requires extensive training) to localize humans in videos. Fig. 16 shows the tracing performance. From the results, we can see for all speed rates, the system finds and keeps tracking the target stably after around 25 frames. We also notice the system can react faster to the high-speed targets, since in these cases the changes of RF phase signals as well as the speed of the target are significant, hence yielding high correlation between the two channels. As a contrast, if the target remains static (the relative speed rate to the sensors is 0), the system can hardly detect the target.

B. Impact of Velocity Threshold

In TagAttention, we use $v_0 = 0.1\ m/s$ as the minimal relative velocity to trigger the matching of the signals from the two channels (Eq. 7). Note this parameter is fixed for all

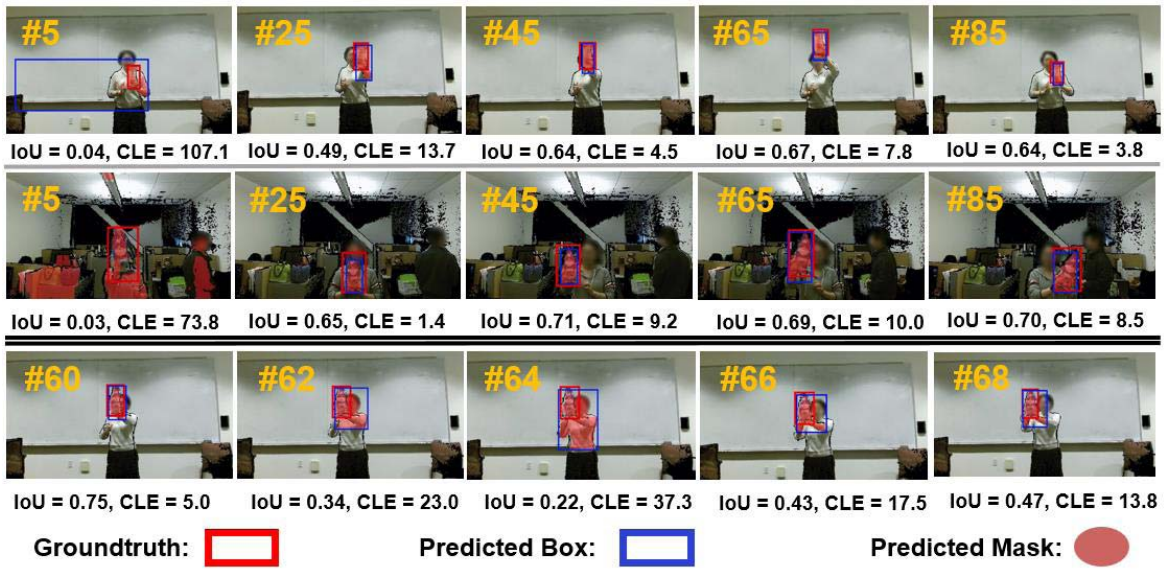
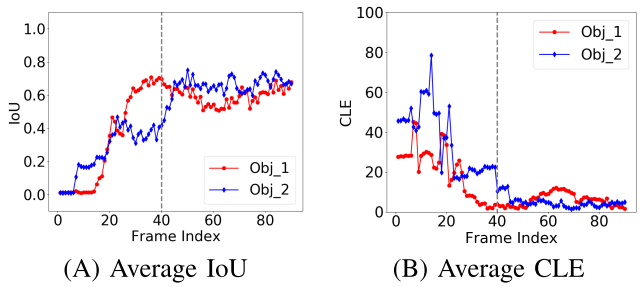


Fig. 13. Examples of single object tracing results.



(A) Average IoU (B) Average CLE

Fig. 14. Tracing results of the two-object scenarios.

above experiments and is empirically selected based on the measurement accuracy of our sensors used in the experiments. In order to illustrate how this parameter impacts the tracing performance, as well as to provide intuition regarding how to set this parameter when using different devices, we present the IoU scores of the tracing system with respect to different v_0 settings in two different scenarios, as shown in Fig. 17.

During the first scenario (the left plot in Fig. 17), the target is static at the beginning and speeds up to around $0.5\ m/s$, while in the second scenario (the right plot), the target speeds from static to around $2.5\ m/s$. From the first scenario, we can see the system with smaller v_0 reacts faster, i.e., the IoU score starts increasing at earlier stage of the motion. At this stage (from frame No. 5 to No. 15), the velocity is relatively small and the system with smaller v_0 is more sensitive to the slow target. However, when the threshold v_0 is set too large (i.e., $0.5\ m/s$), since for the most time of the motion, the target is slower than v_0, the system cannot detect the target at all in those frames. For the second scenario, on the other hand, since the target moves faster than $0.5\ m/s$ for most frames (except the first 15 frames when the target is static), all systems can successfully detect the target within a short latency.

Therefore, in order to reduce the latency of the detection and make use of the maximum number of effective frames during the tracing, v_0 should be set as small as possible. Ideally, with "perfect" sensors that can measure the relative distance accurately in the system, v_0 should be set as 0 to achieve the best performance. However, the commercial sensors we adopt in our experiments can introduce tremendous noise in the sensing data, such as RF phase and depth.

As shown in the first scenario of Fig. 17, although the system with $v_0 = 0.01\ m/s$ reacts slightly faster than the one with $v_0 = 0.1\ m/s$, the overall tracing accuracy is worse than the $v_0 = 0.1\ m/s$ setting. The reason is that for a few certain frames when the target is static or moves slow, the noise in the measurement dominates the changes of RF phase compared with the impact of target velocity. Consequently, for those frames, the attention map computed by Eq. 6 mainly reflects the signal noises rather than the motion of the target. Hence, in practice, we set a larger v_0 (i.e. $0.1\ m/s$) to filter out those frames for the optimal overall performance. As a comparison, in the second scenario, we find the impact of v_0 is much smaller, because the signal noise is neglectable when the target moves fast.

The evaluation result in Fig. 17 suggests us to select v_0 according to the following principles: (1) Choose v_0 as small as possible. (2) v_0 cannot be too small such that the signal noise becomes a significant part of the phase change in the effective frames. (3) It is easier to find an optimal v_0 when the target moves faster. (4) If the application scenario requires to detect slow targets, more accurate sensing devices are necessary.

In addition, we find the following major factors in practice that impacts the velocity measurement accuracy of the RFID reader. (1) manufacture of the RFID reader and tag. (2) signal strength of the antenna, which effects the sampling rate of the target tag. (3) number of the concurrent tags in the environment, which also effects the sampling rate of target. (4) distance of the tag and other dynamic factors in the environment.

C. Impact of Illumination

In this section, we show how the system performs under different illumination conditions. We conduct a comparison

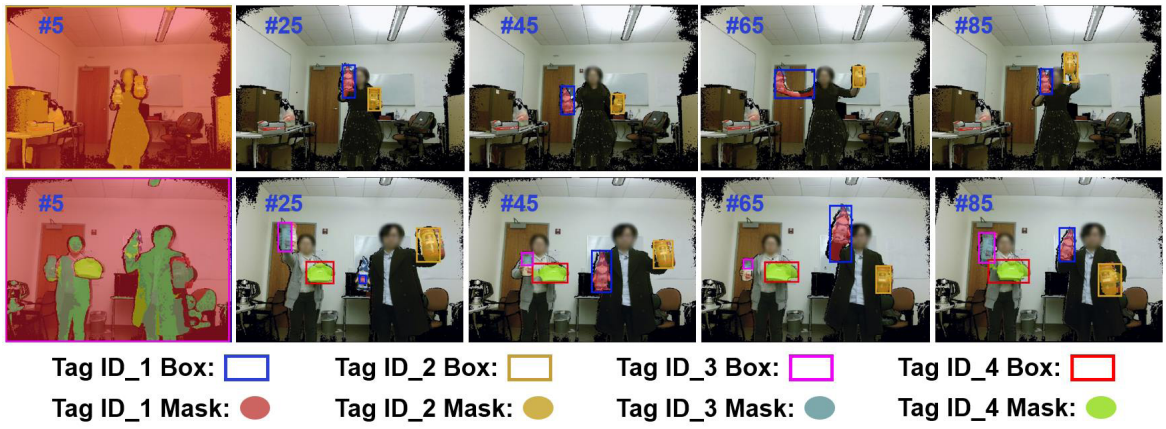


Fig. 15. Examples of multi-object tracing.

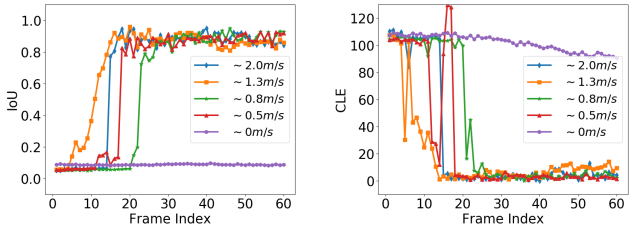


Fig. 16. Tracing performance at different levels of speed rates.

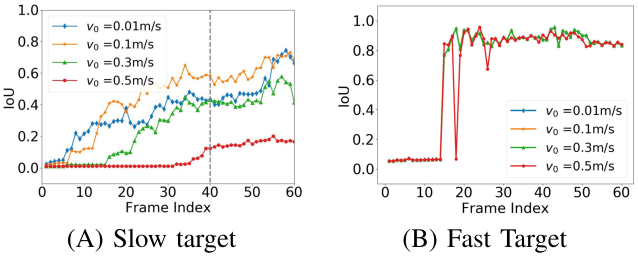


Fig. 17. Impact of parameter v_0.

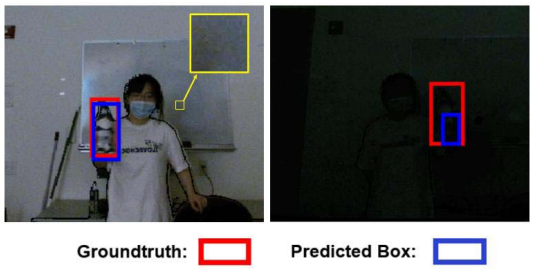


Fig. 18. Moderate (left) and limited (right) illumination scenarios and tracing results. The frames have low SNRs when the illumination is constrained.

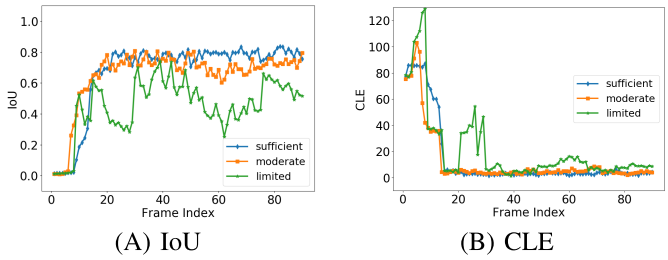


Fig. 19. Impact of illumination.

experiment under three different illumination conditions: sufficient illumination, moderate illumination, and limited illumination. The sufficient illumination scenario is similar as the scenario presented in Fig. 15, while the moderate illumination and limited illumination scenarios are presented in Fig. 18. In the moderate illumination scenario, only some nature lighting through the shutters is allowed in the office. In the limited illumination scenario, the target and the volunteer are almost unseen from the video. With less illumination, the Signal-to-noise ratio (SNR) of the video becomes smaller, which may impact the accuracy of the optical flows learned by the bottom-up attention module.

We repeat the tracing experiment for five times in each scenario, and show the average tracing performance in Fig. 19. From the results, we find the system is robust to the illumination conditions if the target can be seen from the video. However, the performance degrades significantly when the target can hardly be visualized by the camera. In addition, the system is not expected to work when there is no lighting at all.

D. Blockage and Occlusion

Our current system cannot trace the target when the target is blocked due to the following two major challenges. First, the system relies on the visual system to localize the target. If the target is invisible in the video frames due to the blockage, the system cannot acquire the fine-grained localization information of the target. In addition, when the target is blocked, the Non-LoS RF components would dominate the received RF phase signal, which makes it challenging to estimate the distance of the RFID tag.

In Fig. 20, we show how TagAttention would perform in a blockage scenario. In the experiment, we ask a volunteer to use a notebook to block the target toy from being sensed by the camera and RF antenna. From Fig. 20 we find the system can only detect the body part that is exposed to the camera when the target is partially blocked and the RFID tag is fully blocked (frame No. 52 and No. 56). When the target is fully blocked (frame No.60), the system cannot detect the target from the video. Then after the notebook is move away and

Fig. 20. Performance of the system when the target is temporarily blocked.

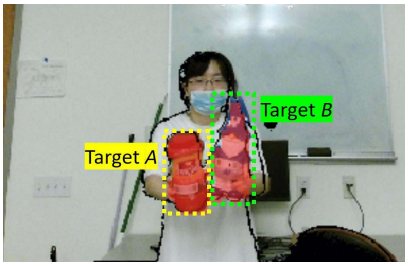


Fig. 21. Two objects A and B move consistently. The dashed boxes represent the groundtruth of the target positions.

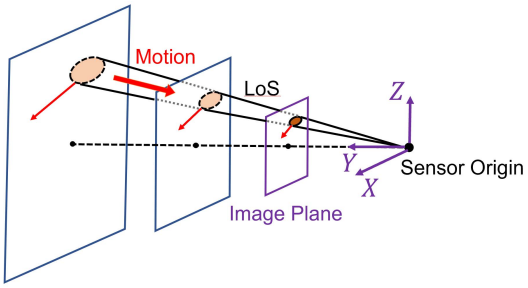


Fig. 22. An illustration of the 2D camera model in a tracing system.

the target exposed to the camera again for a while (around one second), the system is able to capture the target again by the attention mechanism.

There could be two types of possible solutions to resolve tracing with temporary blockage. One is deploying multiple antennas and using advanced RF signal analysis techniques to extract the LoS signal component from the received RF signals [21], [39], [40], [43], then localizing the tag based on the RF signals. However, this type of methods also introduces extra device expenses, sensor calibration and deployment difficulties and larger localization errors. Another type of methods is considering the correlation of visual features of the discovered target over the consecutive frames. For example, since TagAttention can already find the correct mask before the rotation or blockage happens, we may use optical correlation filters [2], [23], which are pretrained on conventional video tracking datasets, to continuously track the targets when they are partially blocked. However, we still cannot correctly localize the target when the target is completely unseen with this type of methods.

E. Consistent and Concurrent Mobile Targets

TagAttention correlates the RFID tag and the target object by the consistency of their motion trajectories. Thus, if two objects move along the same direction and at the same speed to the sensors consistently and concurrently, TagAttention cannot distinguish the two objects. TagAttention has this natural limitation due to the assumption that the system has no prior knowledge about the appearances of the targets to trace and it learns the concept of the "target" from the RFID tags. Hence, the concept of the "target" becomes ambiguous to the system if the two targets move consistently and concurrently.

In Fig. 21, we show how the system would perform in such kind of scenario. In the experiment, we ask the volunteering to hold the two objects (a coffee bottle and a toy) and move the object consistently. In the result, we find the system can highlight the shape of both objects from the video and recognize them as one whole target.

F. Choice of Sensing Technologies

In TagAttention, we utilize an RGB-D camera to capture the fine-grained 3D coordination of video frame pixels. However, RGB-D camera is not a scalable commercial sensing device due to its higher expense ($100 to $200 for most commercial RGB-D cameras) and shorter sensing range than normal RGB cameras. Here we discuss a few other alternative options of the sensing system settings.

Option 1: one RGB camera with one RFID antenna. Though the RGB camera is carefully calibrated (the intrinsic camera matrix is known), we show this system setting is insufficient to trace an unknown object. As the camera model illustrated by Fig. 22, the object appearance projected on the image plane may imply multiple possible positions of the object in 3D space. Any motion vector of an anchor pixel of the target on the RGB image is a projection of the real 3D motion vector on the image plane. We assume we know the "precise" distance from the RFID tag to the sensor by using one RFID antenna in an ideal case (which is impossible in practice due to the 2π wrapping of RFID phase and sensing noises). Then the 3D motion component along the direction of the line-of-sight (LoS) will result in zero position change of the corresponding anchor pixel on the 2D image plane. In addition, without knowing the distance, the 2D motion velocity of the anchor pixel on the image plane can reflect different scales of velocity on the planes that are parallel to the image plane in 3D. Thus, it is infeasible to match a change of RFID signal phase with the projected 2D motion of the tag (or object pixels) in the image plane, without additional information of the RFID tag position or target appearance.

Recent studies adopt this device setting with additional constraints of the target motion space. For example, TagVision [11] and Tagview [12] require the object to move only on a calibrated or fixed 2D subspace.

Option 2: two (or more) RGB cameras with one RFID antenna. A camera stereo system can also provide a depth channel by using multiple calibrated 2D cameras. Thus, the camera stereo system could be another alternative of

the RGB-D sensor, though a commercial stereo camera is no cheaper than an Infrared camera. The channel fusion algorithm of TagAttention can easily be extended to the camera stereo system.

Option 3: one RGB camera with multiple RFID antennas. Another possible solution is to use a single calibrated 2D camera and multiple RFID antennas. The RFID antennas placed at different locations can provide sensing data (such as RSS, signal phase, angle of arrival) from multiple perspectives, which can help reduce the solution space of the projected RFID tag position on the 2D image plane. However, in practice, due to the phase wrapping, multi-path and measurement noise of the RFID sensing signal, two antennas are far from being accurate to obtain a fine-grained projected location, especially when using the commercial RFID readers and antennas as we used in the experiments. Therefore, to make the plan feasible, we need to either increase the number of antennas (such as using antenna arrays) [39], [40], [43] or use software defined radio (SDR) with larger bandwidth [28], or a combination of both, to improve the RFID localization accuracy. Nevertheless, these settings will significantly increase the expense of the sensing devices and difficulties to deploy the sensors in practice.

In summary, we adopt the most commercial sensing hardware setting to achieve the goal of 3D object tracing in TagAtthention. Since vision techniques are more mature and accurate in fine-grained localization of objects, our key insight is to use the vision channel data as the major positioning method of the object and use the RFID channel information to actively detect and identify the unknown visual components from the videos.

VI. Limitations of the Work

Tracing of the mobile target without human's supervision is a critical but challenging problem in wireless sensing and robotics. TagAttention solves detecting and tracking mobile targets with RFID tags in an active manner. Meanwhile, we acknowledge the following limitations of our current work and invite new research ideas to resolve those challenges. First, as discussed in Sec. V-D and Sec. V-E, our current system does not support target tracing when the target is temporarily blocked. Neither can it distinguish the IDs of the objects that move consistently and concurrently. In addition, the system can only detect mobile objects that moves faster than v_0 (Sec. V-B). Due to these limitations, the current system is not ready to be applied in the real-world scenarios that contain too much complex semantics, such as cashier-free stores with tens of tags and customs in a crowded space.

VII. Conclusion

We summarize the contribution of this work as the following:

- We make the first attempt to design a pixel-level RF-Vision fusion system that can detect and track the targets with unknown appearances. The system is mainly based on a novel "attention" model, namely, we use the RF signal as a "top-down" supervision to direct the visual system to discover the target.
- We propose an "attention propagation" method to propagate the per-frame attention maps that contains historical localization information over video frames, so that the system can trace the target in the long-term.
- The system integrates advanced optical flow techniques from Computer Vision and RF signal phase processing from RFID localization studies.
- A calibration-free tracing system is implemented using a commercial RFID reader and an RGB-D camera. We also propose the RF signal smoothing, channel synchronization, and tracking refinement strategies to resolve the practical challenges in the real system.

References

[1] M. Andriluka *et al.*, "PoseTrack: A benchmark for human pose estimation and tracking," in *Proc. IEEE CVPR*, Jun. 2018, pp. 5167–5176.

[2] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE CVPR*, Jun. 2010, pp. 2544–2550.

[3] M. Bouet and A. L. dos Santos, "RFID tags: Positioning principles and localization techniques," in *Proc. 1st IFIP Wireless Days*, Nov. 2008, pp. 1–5.

[4] H. Cai, G. Wang, X. Shi, J. Xie, M. Wang, and C. Qian, "When Tags 'read' each other: Enabling low-cost and convenient tag mutual identification," in *Proc. IEEE ICNP*, Oct. 2019, pp. 1–11.

[5] K. Chawla, C. McFarland, G. Robins, and C. Shope, "Real-time RFID localization using RSS," in *Proc. ICL-GNSS*, Jun. 2013, pp. 1–6.

[6] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "SegFlow: Joint learning for video object segmentation and optical flow," in *Proc. IEEE ICCV*, Oct. 2017, pp. 686–695.

[7] L.-X. Chuo, Z. Luo, D. Sylvester, D. Blaauw, and H.-S. Kim, "RF-echo: A non-line-of-sight indoor localization system using a low-power active RF reflector ASIC tag," in *Proc. ACM MobiCom*, 2017, pp. 222–234.

[8] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current Biol.*, vol. 14, no. 19, pp. R850–R852, Oct. 2004.

[9] H. Ding *et al.*, "Trio: Utilizing tag interference for refined localization of passive RFID," in *Proc. IEEE INFOCOM*, Apr. 2018, pp. 828–836.

[10] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE ICCV*, Dec. 2015, pp. 2758–2766.

[11] C. Duan, X. Rao, L. Yang, and Y. Liu, "Fusing RFID and computer vision for fine-grained object tracking," in *Proc. IEEE INFOCOM*, May 2017, pp. 1–9.

[12] C. Duan, W. Shi, F. Dang, and X. Ding, "Enabling RFID-based tracking for multi-objects with visual aids: A calibration-free solution," in *Proc. IEEE INFOCOM*, Jul. 2020, pp. 1–10.

[13] *EPC Radio-Frequency Identity Protocols Class-1 Generation-2 UHF RFID Protocol for Communications at 860 MHz–960 MHz*, EPCglobal, Brussels, Belgium, 2005.

[14] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," in *Proc. IEEE CVPR*, Oct. 2017, pp. 3038–3046.

[15] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran, "Detect-and-track: Efficient pose estimation in videos," in *Proc. IEEE CVPR*, Jun. 2018, pp. 350–359.

[16] J. Han *et al.*, "CBID: A customer behavior identification system using passive tags," in *Proc. IEEE ICNP*, Oct. 2014, pp. 47–58.

[17] J. Han *et al.*, "Twins: Device-free object tracking using passive tags," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1605–1617, Jun. 2016.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE ICCV*, Oct. 2017, pp. 2961–2969.

[19] J. Impin, "Speedway revolution reader application note—Low level user data support, revision 3.0 2013-09-11," Impinj, Inc., Seattle, WA, USA, Appl. Note 202755318, 2013.

[20] M. Jaimez, C. Kerl, J. Gonzalez-Jimenez, and D. Cremers, "Fast odometry and scene flow from RGB-D cameras based on geometric clustering," in *Proc. IEEE ICRA*, May 2017, pp. 3992–3999.

[21] C. Jiang, Y. He, X. Zheng, and Y. Liu, "Orientation-aware RFID tracking with centimeter-level accuracy," in *Proc. ACM/IEEE IPSN*, Apr. 2018, pp. 290–301.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[23] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE CVPR*, Jun. 2018, pp. 8971–8980.

[24] H. Li, E. Whitmire, A. Mariakakis, V. Chan, A. P. Sample, and S. N. Patel, "IDCam: Precise item identification for AR enhanced object interactions," in *Proc. IEEE RFID*, Apr. 2019, pp. 1–7.

[25] H. Li, P. Zhang, S. Al Moubayed, S. N. Patel, and A. P. Sample, "Id-match: A hybrid computer vision and RFID system for recognizing individuals in groups," in *Proc. ACM CHI*, 2016, pp. 4933–4944.
[26] X. Li, Y. Zhang, and M. G. Amin, "Multifrequency-based range estimation of RFID tags," in *Proc. IEEE RFID*, Apr. 2009, pp. 147–154.
[27] T. Liu, L. Yang, Q. Lin, Y. Guo, and Y. Liu, "Anchor-free backscatter positioning for RFID tags with high accuracy," in *Proc. IEEE INFOCOM*, Apr. 2014, pp. 379–387.
[28] Z. Luo, Q. Zhang, Y. Ma, M. Singh, and F. Adib, "3D backscatter localization for fine-grained robotics," in *Proc. USENIX NSDI*, 2019, pp. 765–782.
[29] Y. Ma, X. Hui, and E. C. Kan, "3D real-time indoor localization via broadband nonlinear backscatter in passive devices with centimeter precision," in *Proc. ACM MobiCom*, 2016, pp. 216–229.
[30] Y. Ma, N. Selby, and F. Adib, "Minding the billions: Ultra-wideband localization for deployed RFID tags," in *Proc. ACM Mobicom*, 2017, pp. 248–260.
[31] R. Mandeljc, S. Kovačič, M. Kristan, and J. Perš, "Tracking by identification using computer vision and radio," *Sensors*, vol. 13, no. 1, pp. 241–273, Dec. 2012.
[32] S. Meister, J. Hur, and S. Roth, "UnFlow: Unsupervised learning of optical flow with a bidirectional census loss," in *Proc. AAAI*, 2018, pp. 1–9.
[33] A. Parr, R. Miesen, and M. Vossiek, "Inverse sar approach for localization of moving RFID tags," in *Proc. IEEE RFID*, Apr. 2013, pp. 104–109.
[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.
[35] L. Shangguan, Z. Li, Z. Yang, M. Li, and Y. Liu, "OTrack: Order tracking for luggage in mobile RFID systems," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 3066–3074.
[36] L. Shangguan, Z. Yang, A. X. Liu, Z. Zhou, and Y. Liu, "Relative localization of RFID tags using spatial-temporal phase profiling," in *Proc. USENIX NSDI*, 2015, pp. 251–263.
[37] X. Shi *et al.*, "TagAttention: Mobile object tracing without object appearance information by vision-RFID fusion," in *Proc. IEEE ICNP*, 2019, pp. 1–11.
[38] D. Sun, E. B. Sudderth, and H. Pfister, "Layered RGBD scene flow estimation," in *Proc. IEEE CVPR*, Jun. 2015, pp. 548–556.
[39] G. Wang *et al.*, "An universal method to combat multipaths for RFID sensing," in *Proc. IEEE INFOCOM*, Jul. 2020, pp. 277–286.
[40] G. Wang *et al.*, "Verifiable smart packaging with passive RFID," in *Proc. ACM UBICOMP*, 2016, pp. 156–166.
[41] G. Wang *et al.*, "HMRL: Relative localization of RFID tags with static devices," in *Proc. IEEE SECON*, Apr. 2017, pp. 1–9.
[42] J. Wang and D. Katabi, "Dude, where's my card?: RFID positioning that works with multipath and non-line of sight," in *Proc. ACM SIGCOMM*, 2013, pp. 51–62.
[43] J. Wang, J. Xiong, H. Jiang, X. Chen, and D. Fang, "D-watch: Embracing 'bad' multipaths for device-free localization with COTS RFID devices," *IEEE/ACM Trans. Netw.*, vol. 25, no. 6, pp. 3559–3572, Dec. 2017.
[44] S. W. Oh, J.-Y. Lee, K. Sunkavalli, and S. J. Kim, "Fast video object segmentation by reference-guided mask propagation," in *Proc. IEEE CVPR*, Jun. 2018, pp. 7376–7385.
[45] L. Xie *et al.*, "TaggedAR: An RFID-based approach for recognition of multiple tagged objects in augmented reality systems," *IEEE Trans. Mobile Comput.*, vol. 18, no. 5, pp. 1188–1202, May 2019.
[46] L. Yang, Y. Chen, X. Li, C. Xiao, M. Li, and Y. Liu, "Tagoram: Real-time tracking of mobile RFID tags to high precision using COTS devices," in *Proc. ACM MobiCom*, 2014, pp. 237–248.
[47] L. Yang, Q. Lin, X. Li, T. Liu, and Y. Liu, "See through walls with COTS RFID system!" in *Proc. ACM MobiCom*, 2015, pp. 487–499.
[48] S. Yoo, K. Yun, J. Y. Choi, K. Yun, and J. Choi, "Action-decision networks for visual tracking with deep reinforcement learning," in *Proc. IEEE CVPR*, Jul. 2017, pp. 2711–2720.
[49] M. Zhao *et al.*, "RF-based 3D skeletons," in *Proc. ACM SIGCOMM*, 2018, pp. 267–281.
[50] J. Zhou and J. Shi, "RFID localization algorithms and applications—A review," *J. Intell. Manuf.*, vol. 20, no. 6, p. 695, 2009.
[51] J. Zhou, H. Zhang, and L. Mo, "Two-dimension localization of passive RFID tags using AOA estimation," in *Proc. IEEE IMTC*, May 2011, pp. 1–5.

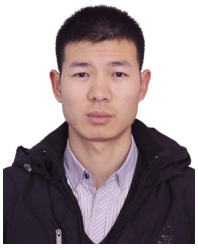
Xiaofeng Shi (Student member, IEEE) received the bachelor's and master's degrees from the Department of Computer Science and Technology, Nanjing University, China, in 2014 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, UC Santa Cruz. His research interests mainly include wireless sensing, learning augmented computer networking systems.

Haofan Cai (Student Member, IEEE) received the B.S. degree from the Southern University of Science and Technology, Shenzhen, China, in 2016. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of California, Santa Cruz. Her research topics mainly focus on RFID and wireless networking.

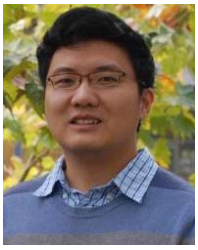
Minmei Wang (Graduate Student Member, IEEE) received the B.E. degree from the Nanjing University of Posts and Telecommunications, in 2014, and the M.Sc. degree from Nanjing University, in 2017. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of California at Santa Cruz. Her research interests include the Internet of Things and network security.

Ge Wang (Member, IEEE) received the B.S. degree from Xi'an Technological University, in 2013, and the Ph.D. degree from Xi'an Jiaotong University. She worked as a Visiting Student with the University of California, Santa Cruz, in 2018 and 2019, respectively. Her research interests include wireless sensor network, RFID, and mobile computing.

Baiwen Huang received the B.S. degree in computer science from the University of California, Santa Cruz. During the summer of his sophomore year, he interned at Turing Robot. Throughout his junior and senior years, he worked in the METX Lab, where he focused on applying machine learning to HIV research, and in Chen Qian's lab at UCSC, where he worked on RFID and computer vision sensing topics.



Junjie Xie (Member, IEEE) received the B.S. degree in computer science and technology from the Beijing Institute of Technology, Beijing, China, in 2013. He was a Visiting Scholar with UC Santa Cruz, from 2017 to 2019. His research interests include distributed systems, software-defined networking, and edge computing.



Chen Qian (Senior Member, IEEE) received the Ph.D. degree from the University of Texas at Austin, in 2013. He is currently an Assistant Professor with the Department of Computer Science and Engineering, University of California Santa Cruz. His research interests include computer networking, data-center networks, software-defined networking, and mobile computing. He is a member of the ACM.