Elevator, Escalator, or Neither? Classifying Conveyor State Using Smartphone Under Arbitrary Pedestrian Behavior

Tianlang He, Zhiqiu Xia, and S.-H. Gary Chan, *Senior Member, IEEE*

***Abstract*—Knowing a pedestrian's *conveyor state* of "elevator," "escalator," or "neither" is fundamental to many applications such as indoor navigation and people flow management. Previous studies on classifying the conveyor state often rely on specially designed body-worn sensors or make strong assumptions on pedestrian behaviors, which greatly strangles their deployability. To overcome this, we study the classification problem under arbitrary pedestrian behaviors using the inertial navigation system (INS) of the commonly available smartphones (including accelerometer, gyroscope, and magnetometer). This problem is challenging, because the INS signals of the conveyor states are entangled by the arbitrary and diverse pedestrian behaviors. We propose ELESON, a novel and lightweight deep-learning approach that uses phone INS to classify a pedestrian to elevator, escalator, or neither. Using causal decomposition and adversarial learning, ELESON extracts the motion and magnetic features of conveyor state independent of pedestrian behavior, based on which it estimates the state confidence by means of an evidential classifier. We curate a large and diverse dataset with 36,420 instances of pedestrians randomly taking elevators and escalators under arbitrary unknown behaviors. Our extensive experiments show that ELESON is robust against pedestrian behavior, achieving a high accuracy of over 0.9 in F1 score, strong confidence discriminability of 0.81 in AUROC (Area Under the Receiver Operating Characteristics), and low computational and memory requirements fit for common smartphone deployment.**

***Index Terms*—Conveyor state classification, smartphone, user behavior, IMU, magnetic field, causal representation learning, evidential model.**

I. Introduction

KNOWING whether a pedestrian is taking an elevator, escalator, or neither is fundamental to many smart city applications. For example, in indoor navigation, such information enhances localization accuracy owing to better detection of floor transition [1], [2], [3], [4]. Such knowledge also plays an important role in understanding pedestrian flow and conveyor preference in a venue, shedding insights on user journey, people management measures, and conveyor capacity planning [5], [6], [7]. However, previous studies on the subjects often employ specially designed body-worn sensors or make strong assumptions on pedestrian behavior, which greatly limits their wide applicability [8], [9], [10], [11].

To overcome this, we study classifying a pedestrian into one of the three *conveyor states* of "elevator," "escalator," and "neither" without any behavior assumption, using the inertial navigation system (INS) commonly available from the off-the-shelf smartphone nowadays. Specifically, we use the multimodal INS readings from the accelerator (namely acceleration), gyroscope (namely angular velocity), and magnetometer (namely magnetic field) to classify the states.[1] Note that we primarily focus on the conveyor states of elevator and escalator in indoor environments; readers interested in transportation modes (such as bus and flight travels) may refer to [14], [15], [16] and the references therein.

Conveyor state classification is challenging, because the measured INS readings are the mixture, or entanglement, of signals due to the two independent processes of conveyor state and arbitrary *pedestrian behavior*. In other words, the underlying process of conveyor state is continuously perturbed by various random and diverse behaviors of pedestrians, including, but not limited to, their spatial movements (such as walking, turning, and accommodating), phone carriage (whether held in hand, stored in pocket, or placed in bag), and various actions (such browsing, swinging, and shaking). These behaviors complicate and perturb the brittle conveyor signals, thereby obscuring the classification decision on conveyor states.

Much effort has been made on using the smartphone INS to classify behaviors regarding human gesture recognition, gait detection, and action recognition [17], [18], [19], [20]. While impressive, their research problems are orthogonal to ours because a pedestrian's conveyor state is determined by the conveyor rather than his/her behaviors.[2] Moreover, these works treat the INS signal as a unit or aggregate in both the training and inference processes. If such a methodology is straightforwardly applied to our case, the accuracy would be unsatisfactory due to

Received 8 January 2025; revised 25 June 2025; accepted 3 July 2025. Date of publication 7 July 2025; date of current version 3 October 2025. This work was supported in part by Research Grants Council Collaborative Research Fund under Grant C1045-23G. Recommended for acceptance by G. Zhu. *(Corresponding author: Tianlang He.)*

Tianlang He and S.-H. Gary Chan are with The Hong Kong University of Science and Technology, Kowloon 999077 Hong Kong (e-mail: theaf@cse.ust.hk; gchan@cse.ust.hk).

Zhiqiu Xia is with Rutgers University, New Brunswick, NJ 08901-8554 USA (e-mail: zx283@scarletmail.rutgers.edu).

Digital Object Identifier 10.1109/TMC.2025.3586618

[1] We forgo barometer due to its relatively lower phone penetration and greater device heterogeneity, leaving it for future study [12], [13].

[2] For example, while people could browse phones on elevators, the browsing behavior cannot be used to define the "elevator" state.

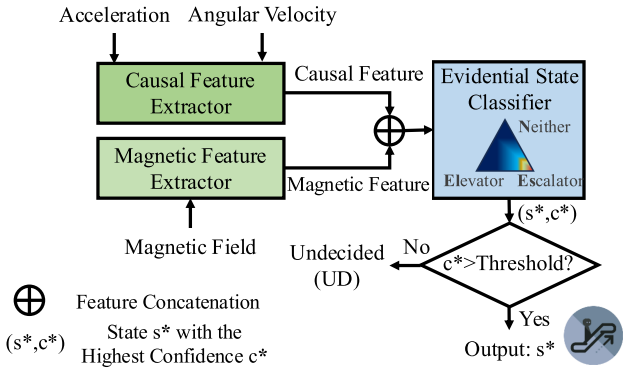


Fig. 1. System overview of ELESON.

the perturbation caused by pedestrian behaviors (as confirmed in our extensive experimental results). Although the general approaches on classification robustness have been applied to INS processing, once extended to our problem, they require precise labels of behaviors during training [21], [22], [23], [24], [25]. This is difficult to implement because exhaustively labeling the plethora of all the possible conceivable behaviors is prohibitively costly, next to impossible. Therefore, classifying the conveyor state under *arbitrary* pedestrian behaviors remains an open and challenging problem.

We propose ELESON, a novel and lightweight deep-learning approach to classify a pedestrian of arbitrary behaviors to **el**evator, **es**calator, **o**r **n**either using the multimodal INS readings from smartphone. We overview ELESON in Fig. 1, which consists of three major modules:

1) *Causal feature extractor to segregate the conveyor motion from pedestrian behaviors:* The motion signals in terms of acceleration and angular velocity capture the movement of elevators and escalators. However, as mentioned before, the motion feature is entangled with pedestrian behaviors. Employing causal decomposition, we propose a causal feature extractor which segregates in deep feature space the motions of moving elevators and escalators, resulting in the causal feature. We design a novel loss function so that such causal feature is extracted independent of pedestrian behaviors.
2) *Magnetic feature extractor to extract conveyor magnetic feature robust against pedestrian behaviors:* The magnetic feature inside the metallic enclosed space of an elevator is different from that of a semi-open escalator. However, some pedestrian behaviors (such as shaking or rotating of phone) often disrupt such magnetic feature. Using adversarial learning, we propose a magnetic feature extractor to capture the conveyor magnetic features robust against these behaviors. Extracting features from the temporal differential of the magnetic field signals, our extractor achieves generalizability for the elevators and escalators unseen in the training data.
3) *Evidential state classifier to estimate the confidence of each state based on the causal and magnetic features:* Applying evidence theory, we employ an evidential state classifier to estimate the confidence of each conveyor state (between 0 and 1) given the causal and magnetic features. The pedestrian is classified to the state with the highest confidence (i.e., s^* and c^* in Fig. 1) if it is above a certain threshold, or "undecided" (UD) otherwise. In contrast to the conventional Softmax-based approaches, our classifier estimates the state confidence that reflects the similarity of the target INS signals with the training data, hence better discriminating the misclassifications.

We curate a large and diverse dataset with 36,420 instances from pedestrians randomly taking elevators and escalators with arbitrary behaviors. We conduct extensive experiments on the dataset, and show that ELESON achieves high accuracy of over 0.9 in F1 score with a strong confidence discriminability of 0.81 in AUROC (Area Under the Receiver Operating Characteristics). As compared with the state-of-the-art classification approaches (all treating the INS signals as a single unit), ELESON outperforms significantly with 14% improvement in F1 score. We have also implemented ELESON on mobile phones and demonstrated that it runs locally in real time with low computational overhead, requiring only 9 MB of memory and consuming less than 2% battery for a 2.5-hour operation.

The remainder of this paper is organized as follows. We first review related work in Section II. Then, we present the problem, the causal feature extractor, and magnetic feature extractor in Section III. After that, we discuss the evidential state classifier to estimate state confidence in Section IV. Finally, we validate ELESON design with extensive experimental results in Section V, and conclude in Section VI.

II. Related Work

Previous studies on classifying the conveyor states often employ specially designed body-worn INS sensors or have specific strong assumptions on pedestrian behaviors [9], [11]. For example, works in [8], [9] study the classification using foot-mounted sensors; works in [10], [11] are based on restricted user behaviors. However, these restrictions could limit their wide deployability. Recently, deep learning has shown powerful capabilities in INS signal processing, effectively classifying various pre-defined behaviors of phone users, including their actions, gestures, and gaits [26], [27], [28], [29], [30], [31], [32], [33]. While impressive, these approaches for human behaviors cannot be satisfactorily extended to the conveyor states of elevator and escalator, because they consider the INS readings as a unit instead of a mixed signal. In our problem, the fragile signals of the underlying conveyor process are frequently perturbed by various arbitrary pedestrian behaviors, which makes it challenging to classify the states robustly. Furthermore, although domain generalization has been applied for robust INS classification, once extended to our problem, it requires precise and exhaustive labeling of pedestrian behaviors in the training process, which is, if not impossible, prohibitively difficult and costly [21], [22], [23], [24], [25], [34], [35], [36]. Therefore, we propose ELESON, the first approach to address arbitrary pedestrian behaviors for conveyor state classification using phone INS without any need for behavior labeling.

Much research work has considered evidential classification for image classification, speech recognition, LiDAR/infrared object detection, etc. [37], [38], [39], [40], [41]. Despite so, the evidential model for INS sensing has rarely been studied. In this paper, we use an evidential state classifier based on the causal and magnetic features of conveyor states and present a loss function for confidence estimation and sound classification.

III. Causal and Magnetic Feature Extraction

In this section, we present the feature extraction of conveyor states under arbitrary pedestrian behaviors. After defining the problem in Section III-A, we discuss the causal feature extractor based on the acceleration and angular velocity in Section III-B, and the magnetic feature extractor based on magnetic field in Section III-C.

A. Problem Definition

A phone-based inertial navigation system (INS) samples the acceleration, angular velocity, and magnetic field in the three dimensions at a fixed interval typically ranging from 1 to 20*ms*. Given a sequence of the signals with T time steps from a pedestrian's phone, or simply an *INS signal*, denoted as $x \in \mathbb{R}^{T\times 9}$, our overarching goal is to classify the *conveyor state* of the pedestrian, denoted as s. Specifically, we define s as a categorical variable that can take one of the three values representing the states of "elevator," "escalator," or "neither."

To achieve this goal, a common practice is training a deep learning classifier that maps an INS signal to the conveyor state given a labeled dataset denoted as $D = \{(x_n, \tilde{s}_n)\}$, where $\tilde{s}_n$ is the label of the nth signal, or simply conveyor state label. After the training process, the classifier is considered ready for testing in real-world scenarios. By doing so, the underlying assumption is that the INS signals used for the training and testing are *independent and identically distributed* (IID), expressed as

$$P\left(x_{test} \mid s\right) = P\left(x_{train} \mid s\right), \tag{1}$$

where x_{test} and x_{train} refer to the INS signals in testing and training scenarios, respectively.

However, the IID assumption may be violated in conveyor state classification due to the *pedestrian behavior*. Specifically, the diverse and arbitrary behaviors of pedestrians often lead to a discrepancy of INS signals in the training and testing, which violates the IID assumption. Formally, under the impact of the conveyor state and pedestrian behavior, the conditional probability of obtaining an INS signal (x) is presented as

$$P\left(x \mid s, V_p\right), \tag{2}$$

where V_p is the variable of pedestrian behavior, and we consider its value drawn from an *uncountable set*.[3] Though in the same conveyor state (s), it is hard to guarantee the pedestrian behaviors (V_p) in testing to match those in training, thus leading to the signal discrepancy, i.e., $P(x_{test} \mid s) \neq P(x_{train} \mid s)$. Such a discrepancy violates the IID assumption, which makes the deep learning classifier unreliable.

To tackle the discrepancy, we apply a feature extraction module before classification. The module aims to extract a *conveyor state feature*, denoted as z, that is statistically independent of pedestrian behavior, shown as

$$P\left(z \mid s, V_p\right) = P(z \mid s). \tag{3}$$

The signal discrepancy hence can be bridged by the extracted features in that

$$P(z_{test} \mid s) = P\left(z_{train} \mid s\right), \tag{4}$$

where z_{test} and z_{train} are the conveyor state features extracted from testing and training scenarios, respectively.

The module has two feature extractors, as shown in Fig. 1. First, we divide an input INS signal into the motion signal (i.e., acceleration and angular velocity), denoted as $x_m \in \mathbb{R}^{T\times 6}$, and magnetic field signal, denoted as $x_b \in \mathbb{R}^{T\times 3}$, shown as

$$x = [x_m, x_b], \tag{5}$$

where $[\cdot,\cdot]$ is the concatenation operation. Then, a causal feature extractor captures a *conveyor causal feature*, denoted as z_c, from the motion signal, and a magnetic feature extractor captures a *conveyor magnetic feature*, denoted as z_b, from the magnetic field signal. Finally, the two features are concatenated to be a conveyor state feature, shown as

$$z = [z_c, z_b]. \tag{6}$$

After the feature extraction, we input the conveyor state feature to the evidential state classifier to be discussed in Section IV. Next, we present the two feature extractors in detail.

B. Causal Feature Extractor

1) Overview: When a pedestrian uses an elevator or escalator, the motion signal on the pedestrian's phone is mainly affected by two *independent* processes: the conveyor transport and pedestrian behavior. The conveyor affects the motion signal due to the transportation process; meanwhile, the pedestrian's various behaviors, whether consciously or unconsciously, more directly influence the phone movement, rotation, pose variation, etc. As a result, the motion signal reflects the mixture of the two processes, and due to various random pedestrian behaviors, the fragile signal of the conveyor transport is difficult to recognize, as illustrated in Fig. 2.

We formulate the generation process of a motion signal under conveyor state (s) and pedestrian behavior (V_p) as a function, denoted as G_m, defined by

$$G_m\left(s, V_p, V_u\right) = x_m, \tag{7}$$

where the variable V_u represents the minor unobserved factors to ensure the rigor of the equation. As mentioned earlier, our goal is to extract a feature of conveyor state that is independent of the pedestrian behavior. However, we neither have the labels

[3]This is because the impact of pedestrian's various behaviors on INS signal (such as their spatial movements, phone carriage styles, actions, and user heterogeneity) is difficult to *precisely* enumerate.

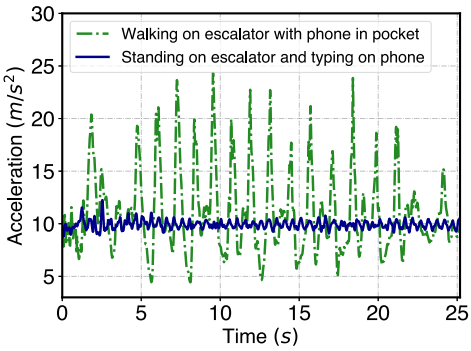


Fig. 2. Illustration of the acceleration signals (in magnitude) in the "escalator" state under different pedestrian behaviors.

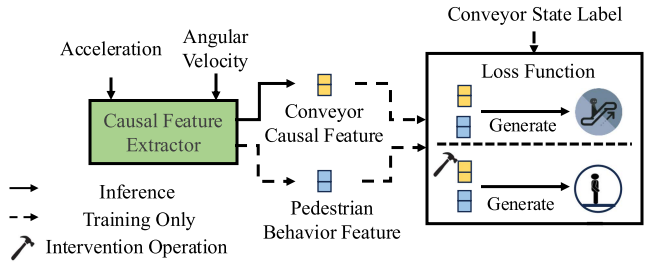


Fig. 3. Causal feature extractor segregates conveyor state from pedestrian behavior based on causal decomposition, learned from a novel loss function. The key idea of the loss function is to simulate an intervention operation in deep feature space.

of pedestrian behavior (V_p) nor the expression of the inverse function of the generation process (G_m^{-1}).[4]

To overcome this, our key idea is to build a deep learning model to conduct a *causal decomposition* on motion signals. Specifically, the model decomposes a motion signal into two deep features that separately encode the conveyor state and pedestrian behavior, such that the feature of conveyor state is independent of pedestrian behavior. As illustrated in Fig. 3, we use a *causal feature extractor*, denoted as f_{θ_m}, to decompose the motion signals (x_m) into a conveyor causal feature (z_c) and a pedestrian behavior feature, denoted as z_p, expressed as

$$f_{\theta_m}(x_m) = [z_c, z_p], \tag{8}$$

where the feature extractor is parameterized by θ_m. In the model training, the causal feature extractor learns from a loss function designed for extracting the *causal feature* of elevator and escalator. In the implementation, the structure of the causal feature extractor is empirically determined, consisting of a two-layer ConvLSTM and two fully connected layers with ReLU as the activation function. We assume that other temporal models should also be suitable [42].

In the following, we discuss the causal feature of elevator and escalator in Section III-B2, and present the loss function in Section III-B3.

2) Causal Feature of Elevator and Escalator: The causal feature of an object reflects the physical property of the object. To name a classic example, temperature generally drops as altitude goes up. Since the law of physics universally holds in most cases, a causal feature is typically more reliable and independent of confounding factors to reflect an object than a common statistical feature [43].

In our problem, we identify that the *signal pattern caused by transportation* is the causal feature of elevator and escalator. To elucidate, as long as an elevator or escalator transports a pedestrian (and his/her smartphone), it inevitably generates the unique transport pattern on the motion signal. Conversely, the transport pattern would not exist without such a transporting process. Therefore, the transport pattern is the causal feature of elevator and escalator. Furthermore, the pattern is naturally *independent* of pedestrian behaviors, because it only depends on the conveyor.

Fundamentally, the transport pattern needs to be extracted from *interventional experiment*. The experiment contrasts pairs of motion signals, between which the conveyor state is treated as univariate, and other variables are strictly controlled. Formally, each *experimental signal*, generated by $G_m(s, V_p, V_u)$, is paired with a *control signal*, generated by $G_m(\text{do}(s), V_p', V_u')$. To ensure the conveyor state (s) as the univariate, we conduct an intervention operation on s, denoted as $\text{do}(s)$, which manipulates the conveyor state to be the "neither" state, and strictly control the other variables such that $V_p' = V_p$ and $V_u' = V_u$.[5] Given the above, the transport pattern, or the causal feature, can be reflected by the difference between the two signals, shown as

$$\Delta x_m = G_m(s, V_p, V_u) - G_m(\text{do}(s), V_p', V_u'). \tag{9}$$

The interventional experiment underlies the causal decomposition upon the experimental signal, which is shown as

$$G_m(s, V_p, V_u) = \Delta x_m + G_m(\text{do}(s), V_p', V_u'). \tag{10}$$

This decomposition is causal because it captures the causal feature of elevator and escalator. In particular, if the conveyor state (s) is the "elevator" or "escalator" state, Δx_m reflects the transport pattern due to the conveyor; if not, Δx_m would be a zero vector, indicating the absence of a conveyor.

With sufficient signal pairs collected from the interventional experiments (or simply *interventional data*), we can learn the causal feature extractor based on (10). However, the interventional data, by and large, are inconvenient to collect due to the demanding univariate setting. In most cases, we only have the experimental signal (or *observational data*) without the control signal. Therefore, we present a loss function for learning the causal decomposition based on observational data.

3) Loss Function: In the below, we simplify the conveyor state as binary in notation, referring to $s = 1$ as either the "elevator" or "escalator" state and $s = 0$ as the "neither" state.[6]

As mentioned earlier, the goal of the causal feature extractor, i.e., $f_{\theta_m}(x_m) = [z_c, z_p]$, is to causally encode the experimental signal, i.e., $G_m(s, V_p, V_u)$. Specifically, z_c encodes the transport pattern, i.e., Δx_m, and z_p encodes the control signal, i.e.,

[4]In other words, handcrafting reliable features of conveyor state under arbitrary pedestrian behaviors could be extremely difficult.

[5]The application of the *do-calculus* follows [43], [44].

[6]This is only for simplifying the equation expressions. The possible values of the conveyor state are still the "elevator", "escalator" and "neither".

$G_m(\text{do}(s), V_p', V_u')$. However, we do not have the control signal and hence the transport pattern as ground truth. To tackle this, we present three learning constraints to enforce the decomposition of the causal feature extractor to comply with (10), based on which we design a loss function.

First, if the decomposition is causal, it should not cause information loss. In (10), the transport pattern and the control signal can be used to reconstruct the experimental signal. Accordingly, we design a loss function that enforces the decomposed features to reconstruct the experimental signal, shown as

$$\mathcal{L}_{rec}(\theta_m, \theta_g) = \sum_{x_m \in D} \text{MSE}\left(g_{\theta_g}(z_c + z_p + \sigma), x_m\right). \quad (11)$$

In the loss function, $\text{MSE}(\cdot,\cdot)$ calculates the mean squared error, $z_c + z_p$ is the vector addition between the two deep feature vectors, σ is a Gaussian noise empirically used to model the V_u in (7), and $g_{\theta_g}(\cdot)$ is a signal generator parameterized by θ_g. In the implementation, the signal generator has three fully connected layers which are trained jointly with the causal feature extractor.

Second, if the decomposition is causal, an intervention operation should effectively remove the causal feature of the conveyor. In the interventional experiment, we carry out an intervention operation to remove the conveyor from the generation process of the experimental signal, which results in the control signal. To simulate this process in deep feature space, combining the conveyor causal feature and pedestrian behavior feature should be able to generate the experimental signal; also, when we remove the conveyor causal feature, the pedestrian behavior alone should generate the control signal. Rewriting this as a constraint gives

$$\begin{aligned} g_{\theta_g}(z_c + z_p + \sigma) &= G_m(s, V_p, V_u), \\ g_{\theta_g}(z_p + \sigma) &= G_m(\text{do}(s), V_p', V_u'), \end{aligned} \quad (12)$$

recalling that $g_{\theta_g}(\cdot)$ is the signal generator. Although the control signal is unavailable, we know that it belongs to the "neither" state. This allows us to give a loose constraint of (12): the pedestrian behavior feature alone should generate the signal whose distribution conforms to the "neither" state, which is shown as

$$g_{\theta_g}(z_p) \sim \mathcal{D}(x_m \mid s = 0), \quad (13)$$

where $\mathcal{D}(x_m \mid s = 0)$ represents the distribution of motion signal in "neither" state. Since the signal is generated from the feature, we directly implement this constraint in deep feature space, facilitated by a classifier denoted as $k_{\theta_k}(\cdot)$. The loss function is shown as

$$\begin{aligned} \mathcal{L}_{sim}(\theta_m, \theta_k) = \sum_{(x_n, \tilde{s}_n) \in D} & \left[\text{CE}\left(k_{\theta_k}(z_c + z_p), s = \tilde{s}_n\right)\right. \\ & \left. + \text{CE}\left(k_{\theta_k}(z_p), s = 0\right)\right], \end{aligned} \quad (14)$$

where $\text{CE}(\cdot)$ is the cross-entropy loss function, and recall that $\tilde{s}_n$ is the conveyor state label. In the implementation, the classifier has two fully connected layers which are trained jointly with the causal feature extractor.

Third, if the decomposition is causal, the control signal should have a larger variance than the transport pattern. In (10), the control signal and the transport pattern are the results of causal decomposition. Since the *diverse* pedestrian behavior only affects the control signal and not the transport pattern, the variance of the control signal should generally be larger than that of the transport pattern. Formally, if the decomposition is causal, we should have

$$\text{Var}\left[g_{\theta_g}(z_c)\right] < \text{Var}\left[g_{\theta_g}(z_p)\right], \quad (15)$$

where $\text{Var}[\cdot]$ calculates the variance of a vector. This constraint could be useful in learning the causal decomposition, as it complements the loose constraint in (13). Similar to the implementation in (14), we directly transform this constraint into a loss function enforced in deep feature space. Specifically, we reduce the variance of the causal feature, and the loss function is presented as

$$\mathcal{L}_{con}(\theta_m) = \sum_{s} \sum_{\substack{(x_n, \tilde{s}_n) \in D, \\ \tilde{s}_n = s}} \text{Var}(z_c \mid \tilde{s}_n). \quad (16)$$

In summary, the loss function of the causal feature extractor is given as

$$\mathcal{L}_{Cal} = \mathcal{L}_{sim} + w_1 \mathcal{L}_{rec} + w_2 \mathcal{L}_{con}. \quad (17)$$

where w_1 and w_2 are the weights for tuning their relative importance. In the implementation, we learn the causal feature extractor using this loss function end-to-end.

Finally, we provide an analysis to interpret the causal feature extractor. Fig. 4 shows an example of the acceleration process of elevator ascending, where the process starts at the 2nd second and ends at the 9th second. The raw acceleration signal is very noisy because the pedestrian perturbs the signal in the process, performing behaviors such as typing, shaking, and moving. This makes the pattern of elevator transport very difficult to recognize. In comparison, the signal generated from the causal feature (aided by the signal generator in (11)) demonstrates the pattern of elevator transport: it first enforces an ascending acceleration, followed by a descending one to maintain a stable speed at the end of the process. This provides an empirical understanding that causal feature enhances the classification of conveyor states.

More rigorously, the causal feature extractor can be regarded as the *importance sampling* on training data to reduce the bias caused by pedestrian behaviors [45]. Applying Bayes' theorem,

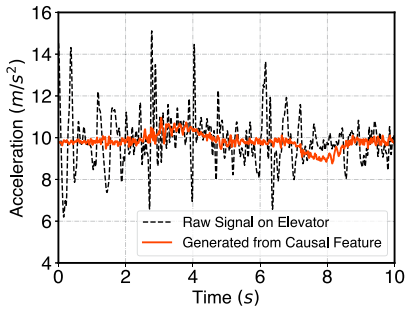


Fig. 4. Illustration of acceleration signal under elevator ascending.

the classification without the loss function can be written as

$$P(s \mid x_m) = \frac{P(x_m \mid s)P(s)}{P(x_m)} = \int P(s \mid V_p)\frac{P(x_m \mid s, V_p)}{P(x_m \mid V_p)}\ dV_p, \quad (18)$$

where the *classification decision* on the left-hand side depends on the distribution of training data on the right-hand side. In other words, the decision depends on not only the motion signal but also the correlation between the conveyor state and pedestrian behavior, i.e., $P(s \mid V_p)$. As mentioned earlier, the conveyor states of pedestrians are largely independent of their behaviors in practice; however, such independence is *not* guaranteed when the training data are observational.[7] As a result, the classification decisions are often *biased* due to the correlation in training data. For example, the classifier may misinterpret the browsing action as a feature of the "escalator" state when browsing frequently coincides with escalators in training data, which may cause errors when a pedestrian browses his/her phone outside an escalator.

To address this, the proposed loss function enforces the conveyor state independent of pedestrian behavior in the training process. This can be interpreted as the importance sampling upon training data distribution. Specifically, the classification decision, after the importance sampling, would only depend on the motion signal, expressed as

$$\begin{aligned} P(s \mid x_m) &= \int wP(s \mid V_p)\frac{P(x_m \mid s, V_p)}{P(x_m \mid V_p)}\ dV_p \\ &= \int \frac{P(s \mid V_p)}{P(s \mid V_p)}\frac{P(x_m \mid s, V_p)}{P(x_m \mid V_p)}\ dV_p \\ &= \int \frac{P(x_m \mid s, V_p)}{P(x_m \mid V_p)}\ dV_p, \end{aligned} \quad (19)$$

where $w = 1/P(s \mid V_p)$ represents the sampling weights for balancing the distribution.

C. Magnetic Feature Extractor

A magnetometer regularly samples the magnetic field orientation at the phone location, generating a sequence of magnetic field signals, or simply a *magnetic signal*, presented as

$$x_b = \left[x_b^{(0)}, x_b^{(1)}, \ldots, x_b^{(T-1)}\right], \quad (20)$$

where the magnetic signal has T samples over time. When a pedestrian uses a conveyor, the signal would vary according to the conveyor movement. Yet another impact is that the signal is sensitive to the metallic structure of the conveyor, either the enclosed shell of the elevator or the semi-open frame of the escalator. Therefore, the magnetic signal may be used to classify the conveyor state, complementing the motion signal, as examplified in Fig. 5.

However, it is not straightforward to classify conveyor states using magnetic signals. Since the magnetic field depends on location, as shown in Fig. 6 (left), the magnetic signals are difficult to be used for the classification in different places. In addition, the signal is affected by many pedestrian behaviors such as waving and rotating phones, which makes the signal noisy in practice.

[7]Note that such spurious correlation does not occur in interventional data.

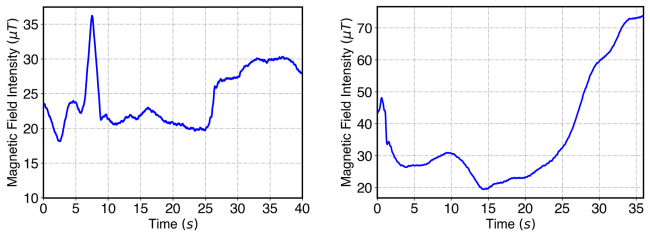


Fig. 5. Examples of magnetic field intensity in elevator (left) and on escalator (right).

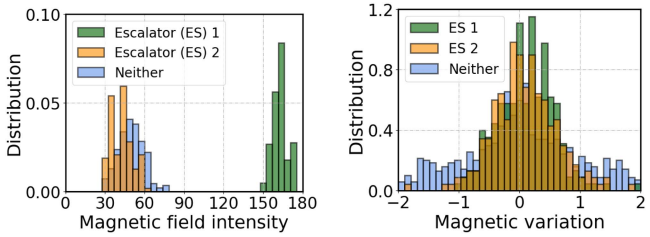


Fig. 6. Evidence that the temporal differential of the magnetic signal (right) exhibits greater stability as a characteristic of the conveyor state compared to raw magnetic field signals (left).

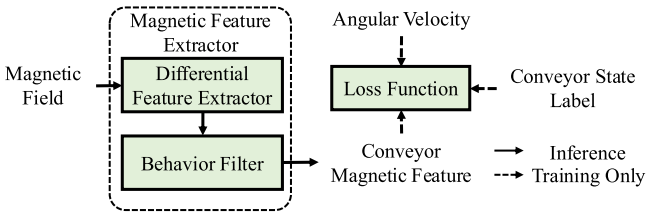


Fig. 7. System diagram of magnetic feature extractor. The differential feature extractor obtains from a magnetic signal a differential feature, based on which the behavior filter enhances the feature robustness against pedestrian behavior. In training, the behavior filter learns from a loss function based on adversarial learning.

To effectively leverage magnetic signals, we aim to extract a conveyor magnetic feature that is independent of conveyor locations and robust to pedestrian behaviors. As shown in Fig. 7, we propose a magnetic feature extractor consisting of a differential feature extractor and a behavior filter. Given a magnetic signal, the differential feature extractor extracts a *differential feature*, denoted as Δx_b, to alleviate the signal dependency on location. After that, a behavior filter reduces the signal noises due to pedestrian behaviors and outputs a conveyor magnetic feature.

The design of the differential feature extractor is based on an empirical finding. As shown in Fig. 6, compared with the raw magnetic signal, the temporal differential of the signal demonstrates much lower location dependency, and it maintains the ability to differentiate the conveyor states. This is because the differential feature of the signal reflects the magnetic variation of the conveyor process, which is much independent of their

locations. Based on this observation, we design a *differential feature extractor*, denoted as $f_b(\cdot)$, which extracts the differential feature by

$$f_b(x_b) = \Delta x_b = \left[\left|x_b^{(t)}\right|_2 - \left|x_b^{(t-1)}\right|_2, t = 1, 2, \ldots, T\right], \quad (21)$$

where $|\cdot|_2$ calculates the magnetic field intensity. Empirically, we find that the intensity is more effective than the orientation for reflecting conveyor states, which has been validated in Fig. 20.

To tackle the pedestrian behavior, we use a *behavior filter* to enhance the feature robustness against pedestrian behaviors. Formally, the behavior filter, denoted as $f_{\theta_b}(\cdot)$, takes a differential feature as input and outputs a conveyor magnetic feature (z_b), defined by

$$f_{\theta_b}(\Delta x_b) = z_b, \quad (22)$$

where the filter is parameterized by θ_b. In the model training, we train the behavior filter robust against the perturbation caused by pedestrian behaviors. Specifically, we regard a behavior as a perturbation when it causes a phone to move or rotate, such as swinging and browsing, denoted as $V_p = 1$, and otherwise, $V_p = 0$. To enhance the robustness, we reduce the discrepancy of the conveyor magnetic feature between the two cases, which is shown as

$$\min \left| P(z_b \mid s = 1, V_p = 0) - P(z_b \mid s = 1, V_p = 1) \right|. \quad (23)$$

To enforce this, we employ a classifier, denoted as $k_{\theta_h}(\cdot)$, to determine the discrepancy between the two distributions. The classifier plays a min-max game with the behavior filter based on adversarial learning [46], and the loss function for the behavior filter is shown as

$$\mathcal{L}_{Mag}(\theta_b) = -\sum_{\substack{(x,\tilde{s}_n)\in D,\\ \tilde{s}_n=1}} \mathrm{CE}\left(k_{\theta_h} \circ f_{\theta_b}(\Delta x_b), V_p\right), \quad (24)$$

where $\circ$ is the composition operator.

In the implementation, we use a threshold of angular velocity ($1.5rad/s$) to determine V_p automatically. This is based on the observation that pedestrians could often lead to a high angular velocity, but elevators and escalators usually cannot. The behavior filter consists of a two-layer ConvLSTM and two fully connected layers with ReLU as the activation function.

IV. Evidential State Classifier

Given the extracted causal and magnetic features, ELESON employs an evidential state classifier to estimate the confidence of the conveyor states. We overview the classifier in Section IV-A, and present its loss function in Section IV-B.

A. Overview

Given a conveyor state feature extracted in a period, say, 2 seconds, we aim to classify the conveyor state and estimate the confidence behind the classification decision. We trust the decision if the confidence is greater than a threshold and remain *undecided* (UD) otherwise. By discarding the decisions with low confidence as UD, we could enhance the system reliability. Therefore, it is important for the confidence to effectively reflect the classification accuracy.

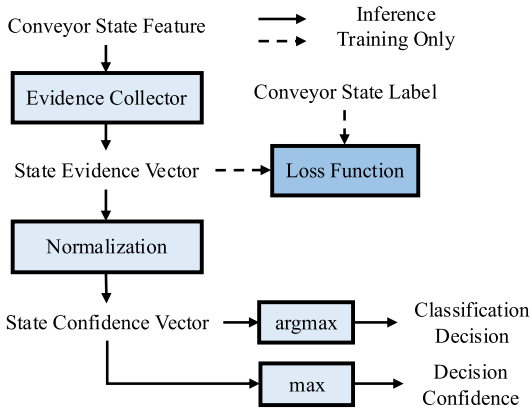


Fig. 8. System diagram of evidential state classifier. Given a conveyor state feature, an evidence collector extracts the evidence supporting each of the states. The evidence is then normalized to be a state confidence vector that gives the classification decision and confidence.

In conveyor state classification, the confidence depends on the INS signal due to the *signal-to-noise ratio* (SNR) and the *signal similarity* to the training data. For example, if pedestrian behavior is much more intense than the conveyor motion, resulting in low SNR, the INS signal could be ambiguous to reflect conveyor states, leading to low confidence. On the other hand, even with a high SNR, an INS signal may not be correctly classified if it is out of the training data of the classifier, which indicates low confidence. To account for the two aspects, existing frameworks are mainly based on the Bayes method and evidence theory [47], [48].[8]

In this paper, we build an *evidential state classifier* due to computational efficiency.[9] The system diagram of the classifier is shown in Fig. 8. Given a conveyor state feature (z), it first uses an *evidence collector*, denoted as $f_{\theta_e}(\cdot)$, to extract a state evidence vector, denoted as E, defined by

$$f_{\theta_e}(z) = E, \quad (25)$$

where $\{e_s \mid e_s \in E, e_s \geq 0\}$ is the evidence value supporting the conveyor state s ($\dim(E) = 3$), and the evidence collector is parameterized by θ_e. Then, it normalizes the evidence vector to be a *state confidence vector*, denoted as C, calculated as

$$C = \frac{E}{e_u + \sum_{e\in E} e}, \quad (26)$$

where $\{c_s \mid c_s \in C, c_s \in [0, 1)\}$ is the confidence of state s, e_u is an uncertainty constant, and we empirically set it to $e_u = \dim(E)$. We select the state with the highest confidence to be the classification decision, which is given as

$$s^* = \arg\max_{c\in C} c. \quad (27)$$

[8]The Softmax-based classifier is often overconfident as it does not consider the signal similarity [47].

[9]Evidential classifier is as efficient as a Softmax-based classifier [49].

Finally, if the highest confidence is larger than a threshold, i.e., $\max(C) > \tau$, we trust the classification decision. Otherwise, the system outputs UD.

In the implementation, the evidence collector consists of three fully connected layers. We use ReLU as both the activation function and the output layer, such that $e_s \geq 0$. The evidence collector is supposed to extract evidence reflecting the two aspects of confidence. In the following, we introduce the loss function for learning the evidence collector.

B. Loss Function

Since the conveyor state of a pedestrian could only be one of the "elevator," "escalator," and "neither", the probabilities of the three states sum up to one. However, considering that the classifier may be unfamiliar with an INS signal (or the input signal is dissimilar to its training data), we use an *uncertainty term* to occupy a fraction of the probability. Specifically, the uncertainty term indicates the "equally likely" among the three states due to the unfamiliarity (namely the epistemic uncertainty [37]). By considering the uncertainty term, the probability assignment depends on the training data of the classifier. Therefore, the probability assigned to each of the conveyor states, i.e., $c_s \in C$, is called the *state confidence*. Formally, the uncertainty term, denoted as u, forms a simplex with the state confidence, which is shown as

$$u + \sum_{c_s \in C} c_s = 1. \tag{28}$$

In evidence theory, confidence comes from *evidence*. The more evidence a classifier collects to support a decision, the less uncertainty it remains. Formally, recall that e_s is the evidence value supporting the state s, and e_u is the uncertainty constant, the uncertainty term can be calculated as

$$u = \frac{e_u}{e_u + \sum_{e \in E} e}. \tag{29}$$

To estimate the evidence value, we learn an evidence collector, and the learning has two objectives.

The first objective is to optimize *classification accuracy*. In other words, the evidence collector should maximize the evidence value supporting the state of ground truth while minimizing the values for the other states. Recalling that x is the INS signal, $\mathrm{CE}(\cdot,\cdot)$ is the cross-entropy loss function, and $\tilde{s}_n$ is the conveyor state label, the loss function of the first objective can be written as

$$\mathcal{L}_{cls}(\theta_m, \theta_b, \theta_e) = \sum_{(x,\tilde{s}_n) \in D} CE\left(\frac{E}{\sum_{e \in E} e}, \tilde{s}_n\right), \tag{30}$$

where the loss function applies to both the feature extraction module and the evidence collector, and this implicitly captures the signal SNR in confidence.

The second objective is to optimize the *uncertainty term*. In other words, the evidence collector should extract more evidence values for its familiar signals, and vice versa. To make the uncertainty term differentiable, we transform the simplex in (28) into a Dirichlet distribution whose variance positively relates to the uncertainty term.[10] Let $a_s = e_s + 1$ and $A = \sum_{e \in E}(e+1)$, the variance of the Dirichlet distribution after the transformation is

$$\mathrm{Var}(E) = \sum_s \frac{a_s(A - a_s)}{A^2(A+1)}. \tag{31}$$

To optimize the uncertainty term, we minimize the variance of the distribution on training data. Overall, the loss function for learning the evidence collector is

$$\mathcal{L}_{Els} = \mathcal{L}_{cls} + \sum_D \mathrm{Var}(E). \tag{32}$$

In summary, recalling that $\mathcal{L}_{Mag}$ and $\mathcal{L}_{Cal}$ are the loss functions of magnetic and causal feature extractors, respectively, the loss function of the whole system is given as

$$\mathcal{L} = \mathcal{L}_{Els} + w_3\mathcal{L}_{Mag} + w_4\mathcal{L}_{Cal}, \tag{33}$$

where w_3 and w_4 are their weights. In the implementation, we learn the classifier and the two feature extractors end-to-end.

V. Illustrative Experimental Results

This section evaluates the performance of ELESON. We first introduce the experimental setting in Section V-A, and present the performance of ELESON in Section V-B. Then, we study system parameters in Section V-C, and show efficiency studies in Section V-D.

A. Experimental Setting

We curate a large and diverse collection of real-world data to validate ELESON. In the data collection, pedestrians freely roam shopping malls with arbitrary behaviors and casually carry their phones, during which they take elevators and escalators at different locations.[11] At the same time, an observer annotate the conveyor state with timestamps whenever the pedestrians get on and off the conveyors (with their consent). In total, we collect data from multiple pedestrians in 10 shopping malls over 20 hours, with roughly 20% in elevators, 20% on escalators, and 60% are neither. To our knowledge, this is the first dataset for classifying the conveyor states of pedestrians under arbitrary behaviors based on phone INS.

Since pedestrians could use conveyors for varied periods, we classify the conveyor state using a short sliding window with size and stride of 2 seconds, leading to 36,420 instances for classification in total. To balance conveyor state labels, we follow object detection and evaluate the performance using F1 score separately for "elevator" and "escalator" states, shown as

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{34}$$

Specifically, $\text{Precision} = \text{TP}/(\text{TP}+\text{FP})$, and $\text{Recall} = \text{TP}/(\text{TP}+\text{FN})$, where TP, FP, FN stand for "true positive", "false positive", and "false negative", respectively. Unless

[10] Proof of equivalence is in [48].

[11] Note that the data are collected from the personal phones of the participants, covering different brands and models.

TABLE I
F1 SCORES UNDER VARIOUS PHONE CARRIAGE (WITH THE CONFIDENCE THRESHOLD OF 0)

Carriage style	Reading	In-pocket	Swing	In-bag
Handcraft	0.73	0.66	0.67	0.60
FootMount	0.75	0.67	0.62	0.66
E2E	0.76	0.76	0.68	0.77
MDG (Rand)	0.79	0.75	0.71	0.73
DIVERSIFY	0.79	0.75	0.72	0.78
MDG (Label)	0.81	0.78	0.73	0.78
ELESON	0.92	0.92	0.84	0.88

stated otherwise, we report the average F1 score over "elevator" and "escalator" states.

To validate ELESON, we compare it with the following state-of-the-art approaches for conveyor state classification:

- *Handcrafted feature approach (Handcraft)* [11] classifies conveyor states using several handcrafted features extracted from phone-based INS signals, assuming a steady holding posture. In the experiment, we use a neural network to classify the extracted features.
- *Foot-mounted sensor approach (FootMount)* [8] classifies conveyor states using a finite-state machine specially designed for foot-mounted INS.

To demonstrate the challenge of our problem, we further compare the performance with the following general classification approaches that have been extended to INS processing:

- *End-to-end Approach (E2E)* [19] is our backbone approach, which classifies INS signal using ConvLSTM end-to-end.
- *Multi-domain Generalization (MDG)* [23] enhances classification robustness using the labels of pedestrian behaviors. To implement this, we label the behaviors using phone carriage styles (Label) and random grouping (Rand), shown in Table I. The implementation is based on ConvLSTM.
- *DIVERSIFY* [21] enhances classification robustness based on implicit labeling, where it employs a clustering algorithm and reduces the feature variance among the clusters. In the implementation, we use ConvLSTM to classify conveyor states incorporating the scheme.

We assess confidence estimation using the area under the receiver operating characteristic curve (AUROC), which measures the discriminability of confidence to distinguish correct and false classification decisions. We regard it as a true positive case when a false decision is regarded as UD. Specifically, AUROC measures the trade-off between false positive rate and true positive rate, which is computed as

$$\text{AUROC} = \int_0^1 S_D(r)dr, \tag{35}$$

where r is the false positive rate and $S_D(r)$ maps r to the true positive rate given a dataset D. We implement the following approaches to compare confidence estimation:

- *Entropy approach (Softmax)* [50] uses information entropy of the classification score to present the confidence of Softmax-based approach;

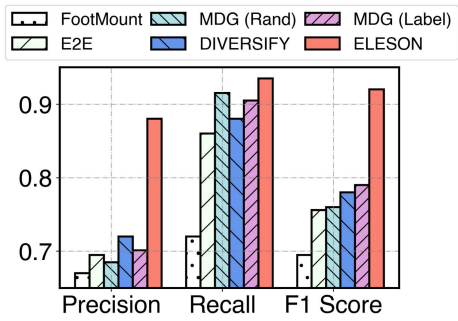


Fig. 9. Under arbitrary pedestrian behaviors, ELESON outperforms previous approaches significantly, improving around 14% in F1 score.

- *Bayesian neural network (Bayesian)* [51] represents confidence using prediction variance over the distribution of network parameters. In the experiment, we sample network parameters using the Dropout mechanism [50];
- *Temperature scaling (TempScale)* [52] calibrates the classification scores using an exponential parameter (calibrated in training) and calculates confidence using the information entropy of the scores.

Furthermore, we have evaluated the computing efficiency of ELESON by deploying it to a mobile phone (Huawei LDN-AL10), based on which we investigated its memory usage, inference time, and power consumption (see Section V-D). In the experiment, each of the conveyor causal feature and the magnetic feature has 128 dimensions, and the model is optimized by the Adam optimizer. The INS sampling frequency is 100*hz*, which is supported by most smartphones [53], [54]. We empirically follow the signal preprocessing in [55], accounting for the general heterogeneity issues, and we assume that other approaches should also be applicable [56], [57]. To reduce randomness, we use the five-fold cross-validation to evaluate the results, where we shuffle the signal sequences of each data collection.

B. Overall Performance

We compare the overall performance of the schemes on the whole dataset in Fig. 9. From the figure, the previous classification approaches for conveyor states fail in our setting due to their behavior or sensor assumptions (note that *Handcraft* performs similarly to *FootMount*). On the other hand, the general classification approaches cannot achieve satisfactory results on our problem, which validates that our problem is new, open, and challenging. In comparison, ELESON has achieved satisfactory precision, recall, and hence F1 score (around 0.89) owing to the feature extraction module. In addition, by setting the confidence threshold to be 0.5, which leads to the UD ratio of 0.05, ELESON gains around 3% improvement in F1 score, while others are less than 2%. With this setting, ELESON has achieved 0.92 in F1 score, improving at least 14% compared to the previous approaches.

Fig. 10 shows the performance of ELESON under different levels of behavior perturbation, where the level is classified by a threshold of angular velocity (1.5 rd/s), and around 40% instances are of the high level. Compared with their performance in the low level of perturbation, all schemes, including ELESON,

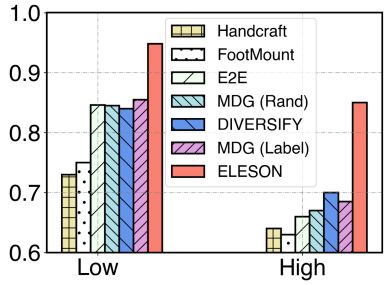


Fig. 10. Under low and high levels of behavior perturbations, ELESON attains a satisfactory F1 score.

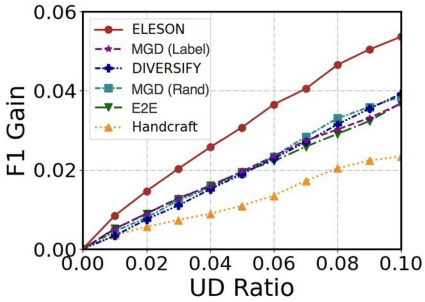


Fig. 11. ELESON benefits higher F1 score from the same UD ratio due to the effective confidence estimation.

degrade under the high level. This is because the behavior perturbation undermines the signal-to-noise ratio. Specifically, the general classification approaches, as they leverage neural networks, perform better than the previous approaches for conveyor state classification; however, their performance is not satisfactory, because they lack sufficiently precise labels of pedestrian behaviors. In comparison, ELESON shows superior F1 scores over 0.85 in both levels without the need for any behavior labels.

Fig. 11 shows how the F1 score varies with the UD ratio, which is tuned by the confidence threshold. In the experiment, the comparison schemes use the Softmax-based classifier as in their original setting. Compared with the other schemes, ELESON gains significant improvements in F1 score under the same UD ratio. This validates that the confidence estimation of ELESON can further enhance the system reliability. Therefore, we set the confidence threshold as 0.5, which leads to 3% improvement in F1 score with merely the UD ratio of 0.05.

Phone carriage style is a coarse but explainable way to label pedestrian behaviors. Table I shows the F1 scores of the comparison schemes under different phone carriage styles with the confidence threshold of 0 (or UD ratio of 0). From the table, all schemes perform relatively better in stable carriage styles (such as reading) than dynamic styles (such as swinging). However, the F1 score of the previous works is less than satisfactory under all phone carriage styles. This validates that they cannot achieve satisfactory results without precise labels of behaviors. In comparison, ELESON achieves a satisfactory F1 score (≥ 0.84) under various carriage styles, which is consistent with the previous results.

Fig. 12 shows an ablation study on the conveyor causal and magnetic features with the confidence threshold of 0. From the

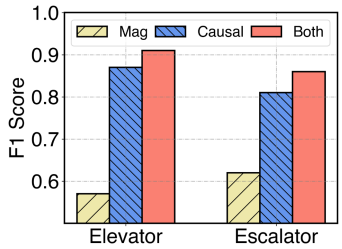


Fig. 12. Ablation study on conveyor causal and magnetic features (with the confidence threshold of 0).

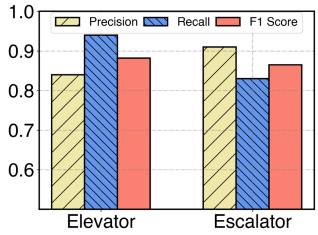


Fig. 13. Performance on unseen elevators and escalators.

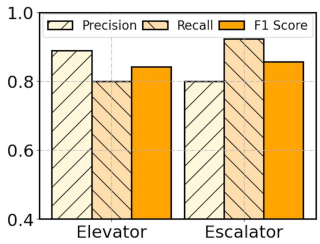


Fig. 14. Performance on crowdsourced data.

figure, the causal feature contributes to the majority of accuracy, and the magnetic feature further boosts the classification effectiveness. This is because they characterize the conveyor state of pedestrians from different aspects. Furthermore, the figure shows that the F1 score of the "elevator" state is higher than that of the "escalator" state. This is because, in our observation, pedestrians in elevators usually perform fewer actions than on escalators, thus introducing less perturbations.

Fig. 13 shows the performance of ELESON on unseen elevators and escalators. In this experiment, we separate training and testing data by shopping malls, such that the conveyors in testing are unseen to the model. In the figure, the F1 score of ELESON slightly reduces by around 3% compared with previous results, due to the subtle shifts of motion patterns and magnetic environments. Despite so, with the enhanced robustness, ELESON achieves satisfactory F1 scores (more than 0.85) on unseen elevators and escalators. This validates the generality of ELESON in practice.

We further evaluate ELESON beyond mall scenarios through a crowdsourcing experiment, and the results are shown in Fig. 14. In the experiment, we collect INS signals from users' daily usage and let the users to mark the time slots (by 15*min*) in which they used a conveyor. Also, the data are from the user devices that are unseen in the training data. We regard it as a true positive if ELESON recognizes the conveyor in a marked

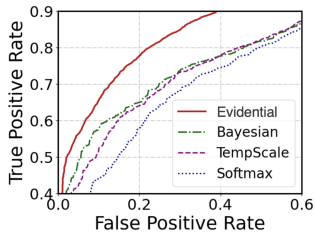


Fig. 15. Study on confidence estimation in AUROC.

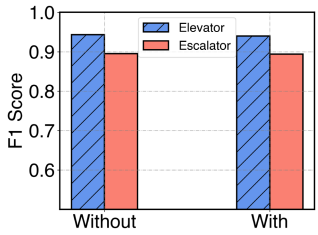


Fig. 16. Ablation study on the uncertainty optimization.

slot or if it remains "neither" state in an unmarked slot. From the figure, ELESON performs similarly to the experimental results in shopping malls, yet the F1 score may slightly slide due to the noisy labels. This validates the generality of ELESON in more real-world scenarios.

Fig. 15 shows the comparison of different confidence estimation approaches on conveyor state classification by ROC curves. The curve indicates a better discriminability of confidence when it is closer to the upper-left corner, and vice versa. In the figure, the entropy-based approach (Softmax) shows poor discriminability due to overconfidence. TempScale universally reduces Softmax confidence, but its improvement in discriminability is limited. While the Bayesian approach can capture epistemic uncertainty, its performance is not stable due to the sampling nature, and more sampling operations lead to heavier computations that are not favorable for the mobile computing. Overall, the evidential state classifier shows strong discriminability of confidence of 0.81 in AUROC with lightweight computation.

Fig. 16 shows the ablation study on the uncertainty optimization in (31). The classifier without uncertainty optimization is labeled by "Without". In the figure, the uncertainty optimization does not influence the F1 score, because it optimizes the evidence collection over all states. This validates that the uncertainty optimization can improve the confidence discriminability without compromising classification effectiveness.

C. System Parameters

In this section, we study the system parameter of ELESON with the confidence threshold of 0.

Fig. 17 shows how the F1 score varies with the weight w_1 in the loss function of causal feature extractor in (17). From the figure, the F1 score increases with the weight when it is less than 0.4 and flats off after that. The gain is because the reconstruction prevents the information loss from decomposition. In the experiment, we use $w_1 = 0.6$.

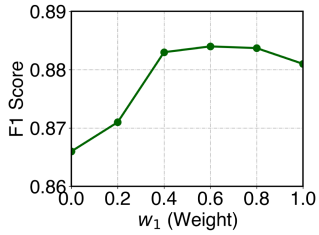


Fig. 17. Parameter study on w_1 in the loss function for causal feature extractor.

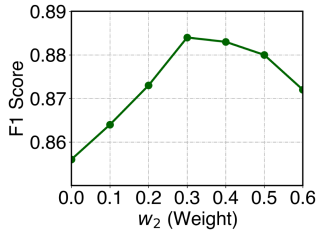


Fig. 18. Parameter study on w_2 in the loss function for causal feature extractor.

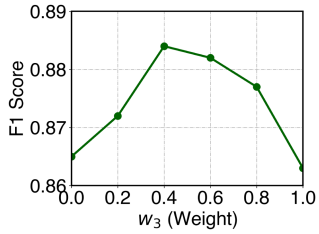


Fig. 19. Parameter study on w_3 for the loss function of magnetic feature extractor.

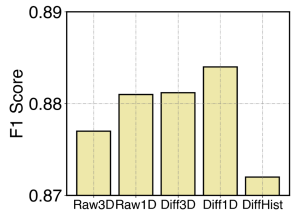


Fig. 20. Ablation study on differential feature extractor.

Fig. 18 plots how the F1 score varies with the weight w_2 in the loss function of causal feature extractor in (17). In the figure, the accuracy shows a U-shape as the weight increases. The F1 score increases because the constraint stabilizes the causal feature. The decrease, on the other hand, is when the weight is so large that the extracted features become inflexible. In the experiment, we use $w_2 = 0.3$.

Fig. 19 shows how the F1 score varies with the weight w_3 in the loss function for the magnetic feature extractor in (33). Similar to Fig. 18, the F1 score shows a U-shape varying with the weight. The increases because the behavior filter enhances the differential feature to be robust against pedestrian behaviors. However, leaning too much on adversarial learning may cause the feature to be inflexible. In the experiment, we use $w_3 = 0.4$.

In Fig. 20, we compare the different implementations for the differential feature extractor in Equation (22). Specifically,

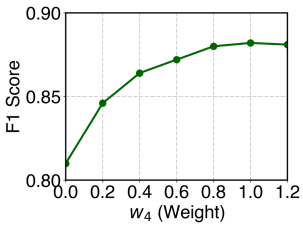


Fig. 21. Parameter study on w_4 for the loss function of causal feature extraction.

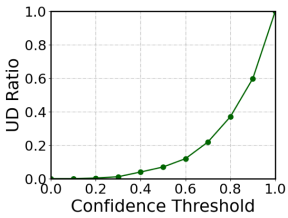


Fig. 22. Distribution of confidence over the whole dataset.

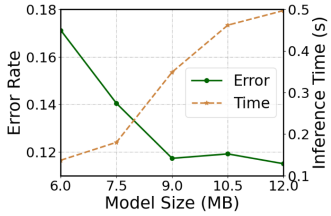


Fig. 23. Inference time and error rate versus model size.

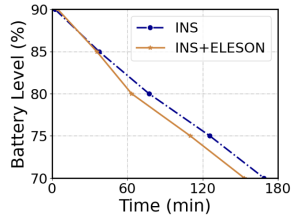


Fig. 24. Study on power consumption.

"Raw3D" stands for the 3D magnetic signal, "Raw1D" is the intensity of the magnetic signals, "Diff3D" is the temporal differential of the magnetic signals, "Diff1D" is the differential of the intensity, and "VarHist" is the histogram on "Diff1D". In the figure, the differential features generally achieve higher F1 score than the raw signals because they are more independent of locations. On the other hand, the intensity of magnetic signals outperforms the 3D orientation as it reduces noise. Finally, the "DiffHist" fails to improve the F1 score as it reduces the important temporal feature of conveyor states.

In Fig. 21, we study how the F1 score varies with the weight w_4 in (33), which is the weight for the loss function of the causal feature extractor. In the figure, the F1 score increases with the weight because the loss function supervises the extraction of the conveyor causal features. In the experiment, we use $w_4 = 1$, where the F1 score flattens off after that.

Fig. 22 shows the distribution of confidence over the whole dataset. The figure shows that the UD ratio grows exponentially with the confidence threshold, with a short tail on the left side. This indicates that ELESON is confident about most decisions. In the experiment, we set the confidence threshold as 0.5, which leads to a UD ratio of 5%.

D. Efficiency Study

In this section, the efficiency studies are conducted on a mobile phone for illustrative purposes, and the conclusion is not limited to the experimental device.

Fig. 23 shows how the error rate and inference time (for making one prediction) vary with the model size. The model size is adjusted by the neuron quantity of each layer, and the error rate equals one minus the F1 score. In the figure, the inference time increases with model size, while the error kneels down at the model size of 9 MB. Therefore, we choose the model size of 9 MB, which is around 0.3% of the phone memory, and each model inference takes around 0.4 seconds. With this setting, ELESON operates in real time (i.e., inference time is less than step size) with a minimal memory budget.

Finally, we show the power consumption of ELESON on a smartphone in Fig. 24. In the experiment, we turn on the INS sensor and use ELESON to process the INS signals. In contrast, we maintain the phone states and turn on the INS sensor without running ELESON. From the figure, INS consumes around 18% of the battery (from 90% to 72%) in 150 minutes, and running ELESON only additionally takes 2% more battery in the same duration. This validates the minimal cost of ELESON in terms of power consumption.

VI. Conclusion

In this paper, we study classifying the conveyor state of a pedestrian with arbitrary behaviors to elevator, escalator, or neither, or simply the conveyor state classification, using the inertial navigation system (INS) on his/her smartphone (i.e., accelerometer, gyroscope, and magnetometer). This research problem is fundamental to many smart city applications, such as indoor navigation and people flow management. The challenge is posed by the arbitrary behaviors of pedestrians, because they entangle with the conveyor states, perturb INS signals, and obscure the classification decision on the states.

We propose ELESON, a novel, effective, and lightweight deep learning approach that classifies the conveyor state under arbitrary pedestrian behaviors using phone INS without the need for any behavior labeling. ELESON separates the motion features of moving elevators and escalators from pedestrian behaviors based on causal decomposition and extracts the magnetic feature of conveyor states based on adversarial learning. Given those features, it uses an evidential classifier to estimate the confidence of each state, which reflects the similarity of an input INS signal to its training data. Through extensive experiments on 36,420 instances of conveyor state data with arbitrary unlabeled pedestrian behaviors collected from ten shopping malls, ELESON shows satisfactory performance, achieving high accuracy of over 0.9 in F1 score and sound confidence discriminability of 0.81 in

AUROC (Area Under the Receiver Operating Characteristics), which improves from previous approaches by 14% in F1 score. Additionally, our efficiency study demonstrates ELESON to operate on a mobile phone in real time (0.4 s for one inference), requiring only 9 MB memory usage and consuming merely 2% battery in 2.5 hours.

ELESON is a pioneering work on classifying the conveyor states of pedestrians using deep learning. In the future, we will extend the scheme to infer the direction of transport or floor transition, and incorporate barometer reading to accomplish richer tasks, higher accuracy, and stronger robustness. We would also like to cover other conveyor types, such as travelators and wheelchairs.

ACKNOWLEDGMENT

The authors would like to thank HONOR for introducing the problem to us and sharing the data for our study.

REFERENCES

[1] G. Wang, D. Zhang, T. Zhang, S. Yang, Q. Sun, and Y. Chen, "Learning domain-invariant model for WiFi-Based indoor localization," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 13898–13913, Dec. 2024.

[2] T. Zhang et al., "RLoc: Towards robust indoor localization by quantifying uncertainty," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 7, no. 4, pp. 1–28, 2024.

[3] W. Zhuo et al., "FIS-ONE: Floor identification system with one label for crowdsourced RF signals," in *Proc. IEEE 43rd Int. Conf. Distrib. Comput. Syst.*, 2023, pp. 418–428.

[4] W. Zhuo et al., "Online path description learning based on IMU signals from IoT devices," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 11889–11906, Dec. 2024.

[5] Z. Cai et al., "Forecasting citywide crowd transition process via convolutional recurrent neural networks," *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 5433–5445, May 2024.

[6] Y. Wang et al., "A coverage-aware high-quality sensing data collection method in mobile crowd sensing," *IEEE Trans. Mobile Comput.*, vol. 24, no. 4, pp. 3025–3040, Apr. 2025.

[7] T. He, J. Tan, and S.-H. G. Chan, "Self-supervised association of Wi-Fi probe requests under MAC address randomization," *IEEE Trans. Mobile Comput.*, vol. 22, no. 12, pp. 7044–7056, Dec. 2023.

[8] N. Kronenwett, S. Qian, K. Mueller, and G. F. Trommer, "Elevator and escalator classification for precise indoor localization," in *Proc. 2018 Int. Conf. Indoor Positioning Indoor Navigation*, 2018, pp. 1–8.

[9] C. Lang and S. Kaiser, "Classifying elevators and escalators in 3D pedestrian indoor navigation using foot-mounted sensors," in *Proc. 2018 Int. Conf. Indoor Positioning Indoor Navigation*, 2018, pp. 1–7.

[10] H. Abdelnasser et al., "SemanticSLAM: Using environment landmarks for unsupervised indoor localization," *IEEE Trans. Mobile Comput.*, vol. 15, no. 7, pp. 1770–1782, Jul. 2016.

[11] H. Liu, R. Li, S. Liu, S. Tian, and J. Du, "SmartCare: Energy-efficient long-term physical activity tracking using smartphones," *Tsinghua Sci. Technol.*, vol. 20, no. 4, pp. 348–363, 2015.

[12] H. Ye, T. Gu, X. Tao, and J. Lu, "SBC: Scalable smartphone barometer calibration through crowdsourcing," in *Proc. 11th Int. Conf. Mobile Ubiquitous Syst.: Comput. Netw. Serv.*, 2014, pp. 60–69.

[13] H. Ye, K. Dong, and T. Gu, "HiMeter: Telling you the height rather than the altitude," *Sensors*, vol. 18, no. 6, p. 1712, 2018.

[14] C. Zhang, Y. Zhu, C. Markos, S. Yu, and J. J. Yu, "Toward crowdsourced transportation mode identification: A semisupervised federated learning approach," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 11868–11882, Jul. 2022.

[15] J. Li, X. Pei, X. Wang, D. Yao, Y. Zhang, and Y. Yue, "Transportation mode identification with GPS trajectory data and GIS information," *Tsinghua Sci. Technol.*, vol. 26, no. 4, pp. 403–416, 2021.

[16] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu, "Transportation mode detection using mobile phones and GIS information," in *Proc. 19th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2011, pp. 54–63.

[17] W. Zhang et al., "mP-Gait: Fine-grained Parkinson's disease Gait impairment assessment with robust feature analysis," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 8, no. 3, pp. 1–31, 2024.

[18] J. Lu and K.-Y. Tong, "Robust single accelerometer-based activity recognition using modified recurrence plot," *IEEE Sensors J.*, vol. 19, no. 15, pp. 6317–6324, Aug. 2019.

[19] Y. A. Andrade-Ambriz, S. Ledesma, M.-A. Ibarra-Manzano, M. I. Oros-Flores, and D.-L. Almanza-Ojeda, "Human activity recognition using temporal convolutional neural network architecture," *Expert Syst. Appl.*, vol. 191, 2022, Art. no. 116287.

[20] H. Xu, P. Zhou, R. Tan, and M. Li, "Practically adopting human activity recognition," in *Proc. 29th Annu. Int. Conf. Mobile Comput. Netw.*, 2023, pp. 1–15.

[21] W. Lu, J. Wang, X. Sun, Y. Chen, and X. Xie, "Out-of-distribution representation learning for time series classification," in *Proc. 11th Int. Conf. Learn. Representations.*, 2023.

[22] J. Yang, Y. Xu, H. Cao, H. Zou, and L. Xie, "Deep learning and transfer learning for device-free human activity recognition: A survey," *J. Automat. Intell.*, vol. 1, no. 1, 2022, Art. no. 100007.

[23] L. Chen, Y. Zhang, Y. Song, A. Van Den Hengel, and L. Liu, "Domain generalization via rationale invariance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 1751–1760.

[24] S. Miao, L. Chen, and R. Hu, "Spatial-temporal masked autoencoder for multi-device wearable human activity recognition," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 7, no. 4, pp. 1–25, 2024.

[25] Y. Kaya and E. K. Topuz, "Human activity recognition from multiple sensors data using deep CNNs," *Multimedia Tools Appl.*, vol. 83, no. 4, pp. 10815–10838, 2024.

[26] D. Zhang et al., "Fine-grained and real-time gesture recognition by using IMU sensors," *IEEE Trans. Mobile Comput.*, vol. 22, no. 4, pp. 2177–2189, Apr. 2023.

[27] H. Prasanth et al., "Wearable sensor-based real-time gait detection: A systematic review," *Sensors*, vol. 21, no. 8, p. 2727, 2021.

[28] X. Chen, S. Jiang, and B. Lo, "Subject-independent slow fall detection with wearable sensors via deep learnin," in *Proc. 2020 IEEE SENSORS*, 2020, pp. 1–4.

[29] Z. Hong et al., "CrossHAR: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 8, no. 2, pp. 1–26, 2024.

[30] M. A. Khatun et al., "Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor," *IEEE J. Transl. Eng. Health Med.*, vol. 10, 2022, Art. no. 2700316.

[31] H. Haresamudram, I. Essa, and T. Plötz, "Contrastive predictive coding for human activity recognition," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 5, no. 2, pp. 1–26, 2021.

[32] G. Saleem, U. I. Bajwa, and R. H. Raza, "Toward human activity recognition: A survey," *Neural Comput. Appl.*, vol. 35, no. 5, pp. 4145–4182, 2023.

[33] F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi, "Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey," *IEEE Access*, vol. 8, pp. 210816–210836, 2020.

[34] N. Bento et al., "Comparing handcrafted features and deep neural representations for domain generalization in human activity recognition," *Sensors*, vol. 22, no. 19, p. 7324, 2022.

[35] T. Shen, I. Di Giulio, and M. Howard, "A probabilistic model of human activity recognition with loose clothing," *Sensors*, vol. 23, no. 10, pp. 4669, 2023.

[36] R. Hu, L. Chen, S. Miao, and X. Tang, "SWL-Adapt: An unsupervised domain adaptation model with sample weight learning for cross-user wearable human activity recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 6012–6020.

[37] W. Bao, Q. Yu, and Y. Kong, "Evidential deep learning for open set action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13349–13358.

[38] C. Wang et al., "Uncertainty estimation for stereo matching based on evidential deep learning," *Pattern Recognit.*, vol. 124, 2022, Art. no. 108498.

[39] A. Pandharipande et al., "Sensing and machine learning for automotive perception: A review," *IEEE Sensors J.*, vol. 23, no. 11, pp. 11097–11115, Jun. 2023.

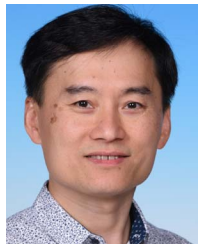
[40] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6700–6713, Oct. 2022.

[41] Y. Xu, T. Bai, W. Yu, S. Chang, P. M. Atkinson, and P. Ghamisi, "AI security for geoscience and remote sensing: Challenges and future trends," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 2, pp. 60–85, Jun. 2023.

[42] Y. Liang et al., "Foundation models for time series analysis: A tutorial and survey," in *Proc. 30th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2024, pp. 6555–6565.
[43] B. Schölkopf et al., "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021.
[44] J. Pearl, "Causal inference in statistics: An overview," *Statist. Surv.*, pp. 96–146, 2009, doi: 10.1214/09-SS057.
[45] S. T. Tokdar and R. E. Kass, "Importance sampling: A review," *Wiley Interdiscipl. Rev.: Comput. Statist.*, vol. 2, no. 1, pp. 54–60, 2010.
[46] D. Lowd and C. Meek, "Adversarial learning," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2005, pp. 641–647.
[47] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," *Artif. Intell. Rev.*, vol. 56, no. Suppl 1, pp. 1513–1589, 2023.
[48] A. Jøsang, *Subjective Logic*, vol. 3. Berlin, Germany: Springer, 2016.
[49] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 3183–3193.
[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[51] P. Thiagarajan, P. Khairnar, and S. Ghosh, "Explanation and use of uncertainty quantified by Bayesian neural network classifiers for breast histopathology images," *IEEE Trans. Med. Imag.*, vol. 41, no. 4, pp. 815–825, Apr. 2022.
[52] A. Karandikar et al., "Soft calibration objectives for neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 29768–29779.
[53] Android, "Android developer documentation," 2025. Accessed: Sep. 05, 2025. [Online]. Available: https://developer.android.com/develop/sensors-and-location/sensors/sensors_motion
[54] Apple, "Apple developer documentation," 2025. Accessed: Sep. 05, 2025. [Online]. Available: https://developer.apple.com/documentation/coremotion/getting-raw-accelerometer-events
[55] S. Herath, H. Yan, and Y. Furukawa, "RoNIN: Robust neural inertial navigation in the wild: Benchmark, evaluations, & new methods," in *Proc. 2020 IEEE Int. Conf. Robot. Automat.*, 2020, pp. 3146–3152.
[56] A. Stisen et al., "Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition," in *Proc. 13th ACM Conf. Embedded Netw. Sensor Syst.*, 2015, pp. 127–140.
[57] X. Ru, N. Gu, H. Shang, and H. Zhang, "MEMS inertial sensor calibration technology: Current status and future trends," *Micromachines*, vol. 13, no. 6, pp. 879, 2022.

Tianlang He received his bachelor of engineering degree (with honor) from Donghua University, Shanghai, China, in 2018, and the master of science degree from the Hong Kong University of Science and Technology (HKUST), Hong Kong, China, in 2019. Currently, he is working toward the PhD degree with the Department of Computer Science and Engineering, HKUST. His research interests include AI Internet of Things (AIoT) and cyber-physical system (CPS), with a focus on system robustness.

Zhiqiu Xia received the bachelor of science degree (with honor) in data science and technology from the Hong Kong University of Science and Technology (HKUST), Hong Kong, China, in 2024. He is currently working toward the PhD degree in electrical and computer engineering with Rutgers University, New Jersey. His research interest includes machine learning and large language model.



S.-H. Gary Chan (Senior Member, IEEE) received the BSE degree (Highest Honor) in electrical engineering from Princeton University, Princeton, New Jersey, with certificates in Applied and Computational Mathematics, Engineering Physics, and Engineering and Management Systems, and the MSE and PhD degrees in electrical engineering with a Minor in Business Administration from Stanford University, Stanford, California. He is currently professor with the Department of Computer Science and Engineering and associate director of GREAT Smart Cities Institute, The Hong Kong University of Science and Technology (HKUST), Hong Kong. He is also board director of Hong Kong Logistics and Supply Chain MultiTech R&D Center (LSCM). His research interests include smart IoT and sensing systems, edge AI, location AI and mobile computing, video/user/data analytics, technology transfer and entrepreneurship. He has been vice-chair of Peer-to-Peer Networking and Communications Technical Sub-Committee of IEEE Comsoc Emerging Technologies Committee, steering committee member and TPC chair of IEEE Consumer Communications and Networking Conference (IEEE CCNC), and area chair of the Multimedia Symposium of IEEE Globecom and IEEE ICC. He has been associate editor of *IEEE Transactions on Multimedia*, guest editor of *ACM Transactions on Multimedia Computing, Communications and Applications*, *IEEE Transactions on Multimedia*, *IEEE Signal Processing Magazine*, *IEEE Communication Magazine*, etc. Through technology transfer and entrepreneurship, He has successfully deployed his research results in industry and co-founded several startups with high commercial and societal impacts. His innovations have received numerous awards and recognitions over the years. Notably, he received Hong Kong Chief Executive's Commendation for Community Service for "outstanding contribution to the fight against COVID-19". He was the recipient of Google Mobile 2014 Award and Silver Award of Boeing Research and Technology. He was a visiting professor or researcher in Microsoft Research, Princeton University, Stanford University, and University of California at Davis. At HKUST, he was director of Entrepreneurship Center, director of Sino Software Research Institute, co-director of Risk Management and Business Intelligence program, and director of Computer Engineering Program. He was a William and Leila fellow with Stanford University, and the recipient of the Charles Ira Young Memorial Tablet and Medal and the POEM Newport Award of Excellence with Princeton University. He is elected fellow of Sigma Xi (FSX) and Chartered fellow of The Chartered Institute of Logistics and Transport (FCILT).