`milliEgo`: Single-chip mmWave Radar Aided Egomotion Estimation via Deep Sensor Fusion

Chris Xiaoxuan Lu[1], Muhamad Risqi U. Saputra[2], Peijun Zhao[2], Yasin Almalioglu[2],
Pedro P. B. de Gusmao[2], Changhao Chen[2], Ke Sun[3], Niki Trigoni[2], Andrew Markham[2]

[1] University of Edinburgh, Edinburgh, Scotland, United Kingdom
[2] University at Oxford, Oxford, England, United Kingdom
[3] University of California San Diego, San Diego, California, USA

ABSTRACT

Robust and accurate trajectory estimation of mobile agents such as people and robots is a key requirement for providing spatial awareness for emerging capabilities such as augmented reality or autonomous interaction. Although currently dominated by optical techniques e.g., visual-inertial odometry, these suffer from challenges with scene illumination or featureless surfaces. As an alternative, we propose `milliEgo`, a novel deep-learning approach to robust egomotion estimation which exploits the capabilities of low-cost mmWave radar. Although mmWave radar has a fundamental advantage over monocular cameras of being metric i.e., providing absolute scale or depth, current single chip solutions have limited and sparse imaging resolution, making existing point-cloud registration techniques brittle. We propose a new architecture that is optimized for solving this challenging pose transformation problem. Secondly, to robustly fuse mmWave pose estimates with additional sensors, e.g. inertial or visual sensors we introduce a mixed attention approach to deep fusion. Through extensive experiments, we demonstrate our proposed system is able to achieve 1.3% 3D error drift and generalizes well to unseen environments. We also show that the neural architecture can be made highly efficient and suitable for real-time embedded applications.

CCS CONCEPTS

• **Computer systems organization** → **Robotic autonomy**; • **Computing methodologies** → *Robotic planning*.

KEYWORDS

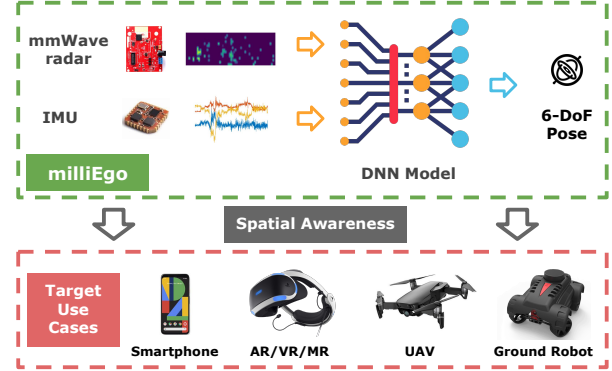Egomotion Estimation; Indoor localization; Millimeter wave radar

Figure 1: Our proposed `milliEgo` uses a low-cost COTS mmWave radar and IMU coupled with a deep neural network model to accurately and robustly estimate 6-DOF egomotion. We envision that `milliEgo` could be widely-integrated on mobile platforms to provide a high degree of spatial awareness to a wide range of applications.

1 INTRODUCTION

From navigating on distant planets to tracking the pose of an augmented reality headset, egomotion awareness plays a vital role in perception and interaction for mobile agents. Unlike map-based localization, egomotion estimation[1] does not require any prior knowledge (e.g., floor plans) about the environment nor any infrastructure setup (e.g., wireless access points), yet it can ubiquitously determine the position and orientation of a mobile agent over time by analyzing sensory data from its movement. Accurate odometry is also a necessary and essential precursor for building mapping, e.g., with SLAM techniques [8, 37].

Owing to their low cost and ubiquity, MEMs inertial sensors (IMUs) have been widely used as the *de-facto* solution to egomotion estimation for a variety of mobile platforms [40, 56]. Unfortunately, the accuracy of MEMs IMUs is limited by noise and bias and consequently inertial odometry suffers from large drift (e.g., SINS [11]) or is impacted by motion dynamics and sensor attachment changes (e.g. PDR [50]). To address the limitations of IMU, multi-modal odometry systems that integrate inertial information with other sensory data (e.g., visual or ranging information) have been proposed. Amongst these system, visual-inertial odometry (VIO) [33] is one of the most dominant, with its advantage of being ubiquitously available in mobile phones etc. However, the performance of

---

[1]Egomotion estimation is also known as odometry estimation in robotics and computer vision community. In this paper, we use *egomotion* to exclusively represent the process of relative pose estimation while *odometry* exclusively refers to the composed trajectory from series of relative poses.

VIO degrades or even fails under adverse lighting conditions (e.g., RGB cameras cannot operate in the dark and depth cameras suffer from glare and strong illumination) [54] or when operating in relatively featureless environments. Similar visibility issues also impact lidar-inertial odometry (LIO), especially in the presence of airborne obscurants (e.g., dust, fog and smoke) [51]. Moreover, lidars are often heavy, bulky and expensive in comparison to cameras and are typically found in high-end robotics rather than in micro-robots or wearable applications.

In order to explore inexpensive alternatives to the above optically based systems, we propose egomotion estimation with single-chip millimetre wave (mmWave) radar in combination with an IMU. Driven by recent advances in the nanoscale CMOS technology [22], these COTS devices have emerged as an innovative low-cost, low-power sensor modality in the automotive industry [26]. A key advantage of mmWave radar over vision is in its robustness to environmental conditions, e.g., it is agnostic to scene illumination and to airborne obscurants. Compared with lidar or mechanically scanning radar (e.g., CTS350-X [67]), single-chip mmWave radars use electronic beamforming and are therefore lightweight and able to fit the payloads of micro robots and form factors of mobile or wearable devices, e.g., the TI IWR6843 even integrates the antennas within the chip package. Smartphones such as the Google Pixel 4 have recently adopted mmWave radar as their on-board sensor for motion sensing [45] while commercial drones have used mmWave radars for obstacle detection [29]. Furthermore, as it is Radio Frequency (RF) based, it does not require optical lenses and can be integrated into plastic housings [27], making them highly resilient to water and dust ingress. We therefore envision that odometry based on a mmWave radar will allow robust egomotion estimation in complex situations (e.g., in subway tunnels, for firefighting and in collapsed buildings), as well as serve as a new enabler for ubiquitous mobility with mobile devices, e.g., for VR/AR.

Transforming this vision into a reliable indoor odometry system, however, requires addressing multiple challenges. *Firstly*, due to specular (mirror-like) reflections, diffraction [41] and significant multi-path, the radar returns are corrupted by noise - our studies indicate as many as 75% of points are outliers. *Secondly*, due to hardware constraints on the number of antennas, the resultant point clouds are highly sparse due to limited angular resolvability, e.g., objects less than fifteen degrees apart are merged into a single point. Thus, a typical radar point cloud has 100x fewer points than a corresponding lidar scan. Such low-quality data makes conventional methods designed for lidar data (e.g., ICP [14]) fail when directly applied to mmWave data. *Thirdly*, although multi-sensor odometry can lower the estimation drift [13], it remains unknown to what extent an mmWave radar can complement inertial and other pervasive sensors e.g. RGB camera, and how best to fuse this. *Lastly*, when applying recent advances in deep neural networks (DNNs) as used in visual or lidar odometry, computational load can be significant which hampers their adoption on mobile, wearable devices and other resource-constrained devices. We investigate whether mmWave radar can be made more computationally efficient due to its inherent sparsity.

Towards addressing these challenges, we propose `milliEgo`, a novel mmWave(-aided) odometry framework that is able to robustly estimate the egomotion of a mobile platform. `milliEgo` follows an end-to-end design and leverages data-driven learning to combat the intrinsic limitations of conventional point registration methods. `milliEgo` demonstrates the feasibility of mmWave odometry and provides a fusion framework to develop a robust mmWave-Inertial Odometry. Despite the use of a deep neural network (DNN), `milliEgo` achieves low latency estimates on embedded platforms. In summary, our contributions are as follows:

- A first-of-its-kind DNN based odometry approach that can estimate the egomotion from the sparse and noisy data returned by a single-chip mmWave radar. Unlike conventional methods relying on explicit point matching, `milliEgo` directly learns the motion transformation, making odometry feasible and reliable.
- Systematically investigating the ideal combination of different neural attention mechanisms for multimodal sensor fusion. Observing the limitation of single-stage self attention, `milliEgo` introduces a two-stage cross-modal attention layer to promote complementary sensor behaviors, yielding robust egomotion estimation (1.3% 3D error drift) in the wild.
- A real-time prototype implementation with extensive real-world evaluations, including testing for both mobile robots and handheld devices. The dataset and code are released to the community.

2 PRIMER

2.1 Principles of mmWave Radar

Range Measurement. The single chip mmWave radar uses a frequency modulated continuous wave (FMCW) approach [60], and has the ability to simultaneously measure both the range and relative radial speed of the target. In FMCW, a radar uses a linear 'chirp' or swept frequency transmission. When receiving the signal reflected by an obstacle, the radar front-end performs a dechirp operation by mixing the received signal with the transmitted signals, which produces an Intermediate Frequency (IF) signal. The distance between object and radar can be calculated from the IF signal [28]. For the radar platform used, the range resolution is ≈4 cm.

Angle Measurement. A mmWave radar estimates the obstacle angle by using a linear receiver antenna array. Signal received at different receiver antennas might have different phase due to the distance between the receivers. Based on the differences in phase of the received signals and the distance between them, the angle of arrival for the reflected signal can be estimated [52]. Although massive MIMO phased antennas are an active area of research [77], they are not commercially available in single chip radars. Instead, the radar used consists of a 3×4 MIMO array, yielding 12 virtual antennas. The resulting angular resolution is poor ($15°$ in azimuth, $58°$ in elevation) and targets which are closely spaced will be 'smeared' together. Subsequent to range and angle estimation, strong peaks are detected which yield a compact set of 3-D points.

2.2 Point Set Registration

At the heart of egomotion estimation from consecutive point clouds is point set registration [43], also known as point matching or point alignment. It aims to find the relative positions and orientations i.e., the transformation of the separately acquired views through maximization of the set of overlapping points. Assume two finite size point sets $\mathcal{A} = \{\mathbf{a}_i\}_{i=1}^{N}$ and $\mathcal{B} = \{\mathbf{b}_j\}_{j=1}^{M}$ acquired by a sensor
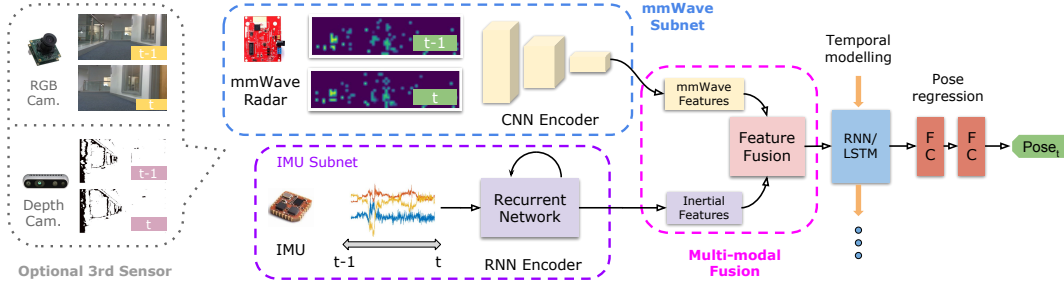
Figure 2: Overview of `milliEgo`, in which *mmWave subnet* and *multi-modal fusion* are two key submodules in its design. The default sensors of `milliEgo` is mmWave and IMU, it can also flexibly integrate the third sensor such as RGB or depth camera.

in a 3D space $\mathbb{R}^3$. In the case $N \leq M$ where $\mathcal{A}$ serves as the reference, a registration starts by searching for correspondence in $\mathcal{B}$. Detected corresponding points are re-ordered in accordance to their counterparts in $\mathcal{A}$, denoted as a new set $\bar{\mathcal{B}} = \{\mathbf{b}_i\}_{i=1}^N$. Once the correspondence is available, a transformation model T on point set $\mathcal{A}$ can be estimated to yield the best alignment between the transformed set $T(\mathcal{A})$ and the corresponding point set $\bar{\mathcal{B}}$. Under the rigid transformation assumption, deriving the optimal T is equivalent to solving a least square minimization problem for all pairs:

$$\min_{\mathbf{R} \in SO(3), \mathbf{t} \in \mathbb{R}^3} \sum_{\mathbf{a}_i \in T(\mathcal{A})} \sum_{\mathbf{b}_j \in \bar{\mathcal{B}}} (\mathbf{b}_j - \mathbf{R}\mathbf{a}_i - \mathbf{t})^2 \qquad (1)$$

where $\mathbf{R} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^3$ are the unknown rotation and translation in model T, from which the egomotion is determined. Intuitively, the precision of point set registration heavily relies on the accuracy of point association. If association is noisy, the subsequent transformation will be inaccurate which will cause egomotion estimation to rapidly degrade over time.

3 OVERVIEW

3.1 Problem Statement

In this work, we consider the general problem of multi-modal odometry, taking as inputs raw data from different types of sensors. An arbitrary sensor input is represented as $\mathcal{X}_f = \{\mathbf{x}_i\}_{i=1}^K$, where K represents the size/length of the data points. In the case of `milliEgo`, we consider a collection $\mathcal{X} = \{\mathcal{X}_M, \mathcal{X}_I\}$ where the data are collected by a mmWave radar (M) and an inertial sensor (I). However, note that our framework is able to directly generalize to a three sensor case, e.g. $\mathcal{X} = \{\mathcal{X}_M, \mathcal{X}_I, \mathcal{X}_V\}$ when a camera (V) is also available. The underlying relative translation $\mathbf{t}$ and rotation $\mathbf{r}$ between a pair of frames is observed by a moving platform[2]. The system goal is to estimate the platform's 6-DoF egomotion $\mathbf{y} = [\mathbf{t}, \mathbf{r}]$. This estimation process is equivalent to learning an effective mapping $\mathbf{x} \rightarrow f(\mathbf{x})$ so as to minimize the error between the prediction $f(\mathbf{x})$ and the ground truth $\mathbf{y}$. Notably, the output of the egomotion model is the 6-DoF relative pose, computed over a pair of consecutive samples. For this reason, we follow [63] and leverage a SE(3) composer to incrementally stitch a series of relative poses into global *odometry*

poses, i.e., the 6-DoF trajectory. As stated in §8.1, the performance evaluation will be mainly based on the trajectory tracking.

3.2 System Overview

From a system-level perspective, `milliEgo` consists of multi-modal mmWave radar and inertial sensing which is then input to algorithms for egomotion estimation; this multi-modal framework is flexible to extend a third sensor in fusion as shown in Fig. 2. The main contribution of `milliEgo` is the end-to-end trainable deep learning model to estimate the platform egomotion. Two novel submodules are critical to the model efficacy that address different design challenges respectively: (i) the mmWave Subnet and (ii) multi-modal fusion module. In what follows, we will first describe the overall network architecture of the proposed `milliEgo` in §4.

4 NEURAL NETWORK ARCHITECTURE

As shown in Fig. 2, the proposed DNN egomotion model starts with multiple egomotion subnets coming from different sensor sets. These multi-modal egomotion features are then fused by a specific fusion module in the DNN. A recurrent neural network models the temporal dependency of the fused features and forward it to fully-connected layers. These layers learn to regress the final features to a pose supervised by the loss function of 6-DoF estimation.

For readability, we focus on the basic DNN architecture in this section, before introducing the *mmWave Subnet* in §5 and the *multi-modal fusion module* in §6 .

Multi-modal Egomotion Subnets. Besides our proposed mmWave subnets, the DNN model also has egomotion subnets from IMU and other sensors depending on their availability (see Fig. 2). In particular, we use the IONET structure [11] to develop the egomotion subnet for IMU and DeepVO [62] structure to develop subnets for RGB and depth cameras. These subnets output their egomotion features to the mixed-attention sensor fusion block.

RNN Temporal Dependency Modeling. On receiving the fused multi-modal features, an RNN is used to model the long-term dynamics. A prominent advantage of using an RNN for egomotion estimation lies in its superior ability in maintaining the memory of hidden states over long term even if the given inputs are noisy. The issue associated with RNNs, however, is that they cannot directly model high-dimensional raw sensory data such as images. But thanks to the CNN extractors, the data fed to RNN are already compressed in a low-dimensionality space. In particular, the RNN used is a two-layer Long Short-Term Memory (LSTM), each of

---

[2]For the ease of model learning and fast convergence, we follow the practice of [54] and represent the rotation $\mathbf{r}$ as an Euler angle rather than a rotation matrix $\mathbf{R}$ as discussed in §2.2.
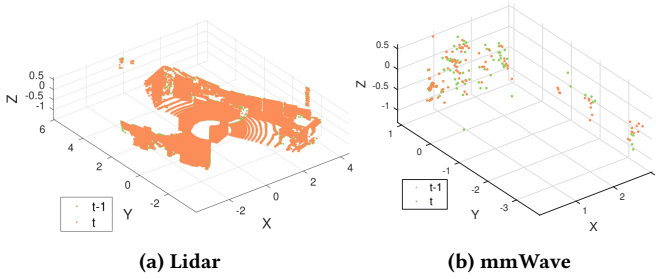
(a) Lidar   (b) mmWave

Figure 3: Comparison between registering lidar and mmWave point clouds (captured at the same instants) with ICP, a commonly used technique. The effectiveness of registration is evidenced by the level of overlap between two consecutive point clouds, shown in orange and green. (a) shows how the dense lidar data can be robustly aligned. In contrast (b) shows that the sparse and noisy mmWave radar point clouds cannot be robustly aligned due to insufficient correspondence. This figure is best viewed by zooming in.

which contains 512 hidden units. This RNN module is on top of the CNN feature extractor layer.

FC Pose Regressor. Given the outputs of RNNs, three fully connected (FC) layers are introduced to regress these sequential features and predict relative transformation. These FC layers have 128, 64 and 6 units respectively, with the last layer regressing features into a 6-DoF relative pose estimate: $\hat{\mathbf{y}} = [\hat{\mathbf{t}}, \hat{\mathbf{r}}]$, where $\hat{\mathbf{y}} \in \mathbb{R}^6$. To promote model generalization ability, we also use a dropout [16] rate of 0.25 between FC layers to help regularization.

Loss Function. Putting it together, deriving the parameters of our network model requires minimizing the following Mean Squared Error (MSE) loss:

$$\mathcal{L} = \frac{1}{K} \sum_{k=1}^{K} ||(\hat{\mathbf{t}}_k - \mathbf{t}_k)||_2^2 + \gamma ||(\hat{\mathbf{r}}_k - \mathbf{r}_k)||_2^2 \qquad (2)$$

where $||\cdot||$ denotes $L2$-norm and γ is a hyper-parameter to balance the weights of translation and rotation errors. Training details of the model are provided in §7.

5 DEEP MMWAVE EGOMOTION SUBNET

Although deep models for inertial and visual egomotion [11, 15, 62] are well established and can be reused, mmWave based egomotion remains an open area and to the best of our knowledge has not been considered from a data-driven perspective. As such, we will first elaborate on the technical challenges created by the single-chip mmWave radar and then introduce our end-to-end solution.

5.1 Challenge: Noisy Correspondence

The fundamental modeling challenge associated with the use of single-chip radar is with their noisy and sparse point clouds e.g. containing tens of points rather than the thousands as typically acquired by lidar or depth sensors. As a result it is very challenging finding reliable transformations between consecutive frames when using traditional point matching methods [21].

5.1.1 Why does the challenge arise? As described in Sec. 2.2, the quality of point set registration between consecutive point clouds

underpins accurate egomotion estimation. Critical to most registration techniques is the determination of correspondence between spatially localized 3D points within each cloud. Such a registration problem is hard to solve when facing issues of noise, missing or spurious data [21], all of which are characteristic of the point cloud generated by the single-chip mmWave radar. In particular, the two main limitations are (1) multi-path effects and (2) sparsity, both of which are discussed as follows.

Multi-path Noise: Multi-path noise is a longstanding issue for almost all RF technologies. Signal propagation of mmWave radar in indoor environments is subject to multi-path [71] due to beam spreading, diffraction and reflection from surrounding objects. As a result, reflected signals arriving at a receiver antenna are often from two or more paths, incurring smearing and jitter. the primary contributor to the non-negligible proportion of noise artefacts and spurious data in the mmWave point cloud.

Sparsity: On top of the strong noise, mmWave point clouds also have severe sparsity issue, owing to (1) the fundamental *specularity* of mmWave signals and (2) the low-cost single-chip design. Wireless mmWave signals are highly specular i.e., the signals exhibit mirror-like reflections from objects [39]. As a result, not all reflections from the object propagate back to the mmWave receiver and major parts of the reflecting objects do not appear in the point cloud. Moreover, as opposed to massive MIMO radar technologies, the mmWave radar in this case only has 3×4 antennas which is effective in both cost and size but greatly limits the ability to resolve spatially close targets. Moreover, in order to lower bandwidth and improve signal-to-noise ratio, algorithms such as CFAR (Constant False Alarm Rate) [66] are used for data processing and *only* provide an aggregated point cloud, further reducing density. For these reasons, the resulting point clouds only contain approximately ∼ 100 reflective points per frame which is over 100× sparser than a lidar.

5.1.2 The failure of traditional methods. The multi-path noise and sparsity issues discussed above collectively impact the performance of conventional geometry based methods. Voting-based registration methods are rendered ineffective [25] through extreme sparsity. The other established class is the Iterated Closest Point algorithm (ICP) and its variants, achieving greater robustness by iteratively interleaving correspondences matching and transformation estimation. Nonetheless, ICP methods require a good initial estimate which is typically achieved with Random sample consensus (RANSAC) [24] and other bootstrapping methods. Again, due to the sparsity and lack of resolution for feature extraction, these algorithms perform poorly, exacerbating convergence to local rather than global minima [6]. Fig. 3 compares the performance of using RANSAC-ICP algorithm to register lidar and mmWave point clouds, both captured at the same time from the same view. Firstly, it is immediately apparent how much denser the lidar point cloud is, accurately representing 3-D scene structure. Secondly, although the lidar point clouds are accurately registered, the mmWave radar point clouds cannot be accurately registered as seen by the lack of alignment. Experimentally, we observe that pose estimation error is ∼ 9-fold higher for mmWave radar compared with lidar. The direct consequence of such poor and ineffective registration is degraded egomotion estimation, rapidly causing the global trajectory to drift.

5.2 mmWave Sub-network

Motivation. Addressing the challenge of noisy correspondences requires more sophisticated estimation algorithms beyond simple heuristics. We thus propose to use data-driven deep learning based approaches which have proven to be effective in extracting useful motion information from complex sensor observations. Critical to this design choice is that deep learning approaches allow for end-to-end modeling and treat the challenging mmWave egomotion problem as a subnetwork or a branch in the full neural network model. Our key insight is that instead of a traditional two-step registration process, it is possible to use a deep neural network that jointly optimizes the transformation estimation and implicit correspondence association between consecutive frames. This end-to-end learning approach helps the system effectively counter the above challenges. Therefore, the knock-on effect between the inter-leaving correspondence association and transformation estimation is mitigated. Furthermore, the end-to-end design also allows us to readily integrate odometry branches from other sensors, since they are treated as other subnets in the full network.

mmWave Subnet Inputs. Determining an effective input format to the mmwave subnet is non-trivial. Point cloud data are unordered and irregularly sampled in 3D space. Extracting the structured motion patterns from such unstructured data is already difficult, not to mention the issues with ambiguous ordering due to severe sparsity and noise. This is likely the reason why existing networks designed for motion flow extraction often require structured inputs (e.g., [18, 76]). To combat the impact of lack of data structure and irregularity, we encode mmWave point clouds into 2D panoramic-view images, following the convention of [34]:

$$
\begin{aligned}
\alpha &= arctan2(p_y, p_x), \quad r = \lfloor \alpha/\Delta\alpha \rfloor \\
\beta &= arcsin(p_z/\sqrt{p_x^2 + p_y^2 + p_z^2}), \quad c = \lfloor \beta/\Delta\beta \rfloor
\end{aligned}
\tag{3}
$$

where (p_x, p_y, p_z) denotes a 3D point and (r, c) is the 2D map position of its projection. α and β denote the azimuth and elevation angle when observing the point. $\Delta\alpha$ and $\Delta\beta$ are the average horizontal and vertical angle resolution between consecutive beam emitters respectively. After panoramic projection, we further normalize values to the range $[0, 255]$. Points closer to the sensor are assigned higher values. Through this transformation, unstructured 3D point clouds are encoded to structured 2D *'depth'* images, in a similar spirit to RGB images which are tractable for many established DNN feature extractors. One key observation we found in this work is that, with right DNN, meaningful egomotion features can be learned from such *'depth'-like* mmWave input representations.

CNN Feature Extractor. Given the subnet inputs, the next step is to design a feature extractor to obtain useful mmWave features for motion estimation. As the input mmWave data have been converted into the format of images at this point, we adopt a CNN structure to process it and extract egomotion features. Notably this design rationale is in a similar spirit of [18], in which implicit correspondences (or the optical flow) between a pair of RGB images was estimated by a CNN. Taking as inputs two consecutive panoramic images $(\mathbf{x}_{k-1}^M$ and $\mathbf{x}_k^M)$, our extractor comprises 9 convolutional layers and each layer is followed by a Leaky version of a Rectified Linear Unit (LeakyReLU) non-linearity activation. The sizes of the receptive fields in the network gradually reduce from 7×7 to 5×5 and then
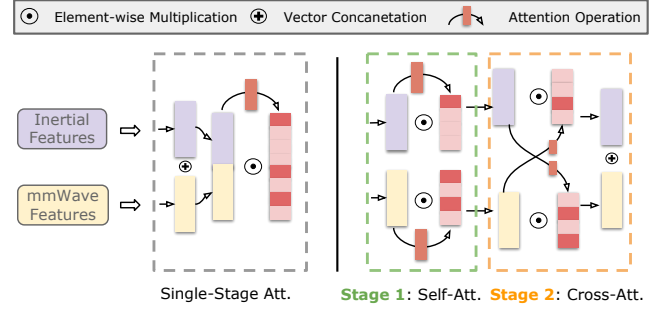

Figure 4: Architectural comparison between single-stage attention mechanism (e.g. [13, 54]) and our proposed mixed (two-stage) attention mechanism.

3×3 to incrementally capture smaller features. The extracted latent mmWave feature representation not only compresses the original high-dimensionality panoramic data into a compact description, but facilitates the temporal dependency modeling. When there is only one mmWave radar in `milliEgo`, the extracted features from the mmWave subnet are used as a branch to be fused with other modalities. However, we will demonstrate the effectiveness of this architecture in enabling mmWave-only odometry in §8.2.4, where multiple mmWave radars are co-located on a mobile platform.

6 MULTI-MODAL EGOMOTION FUSION

So far we have introduced the egomotion estimation subnets of mmWave radar and other sensors. The question remains is how to fuse the multi-modal egomotion features such that the model can adaptively cope with different situations by using the 'right sensors' at the right time.

6.1 Challenge of Adaptive Fusion

A robust multi-modal egomotion estimation is based on the premise of complementary interactions amongst different sensors, whose roles are adaptively altered or re-weighted in response to observation uncertainties or self/environmental dynamics. For example, the mmWave sensor should dominate egomotion estimation when the IMU is degraded in performance. Although classical fusion methods can realize such an adaptation by incorporating physical models into the algorithm design (e.g., Bayesian filtering or fixed-lag smoothers) they require expert design. However, adaptive fusion becomes much harder in the deep feature space. This challenge is rooted in the long-standing criticism of DNN: it is difficult for one to manually incorporate an adaptation mechanism into a black box model due to the complex multilayer nonlinear structure in DNN [2].

6.2 Mixed Attention for Fusion

To endow a multimodal deep egomotion network with the ability to adapt to environmental uncertainties, we systematically investigate the combination of different attention mechanisms in order to best exploit complementary sensor behaviors. The intuition behind attention in general is that features are not always of equal importance and their relative contribution varies on the context.

Critical to this module is to study the effectiveness of a mixed or two-stage attention mechanism. As shown in Fig. 4, mixed attention consists of two attention layers corresponding to the intra- and inter- sensor behavior regulator in our fusion framework. Our goal is to investigate whether the mixed attention framework can better address the fusion challenge than single-stage attention method and effectively learn how best to exploit the complementary nature of the individual sensors. In what follows, we introduce the details of the *self-attention* module for individual sensors, and the *cross-modal attention* module which deals with interactions across sensor modalities.

6.3 Intra-sensor Self-Attention

A promising way to enable model adaptation is by letting individual odometry branches adapt themselves first, similar to a self-filtering process that autonomously sieves informative features. This concept of 'self-attention' was initially designed for machine translation and image transformation to address the long dependency issue [19, 47, 61]. Inspired by these early successes, we consider using the self-attention mechanism to realize egomotion self-regulation. Without resorting to the complex query-key-value pipeline in [61], we adopt a space/spacetime non-local framework akin to [65] which is proven to adaptively reweight feature importance.

For ease of illustration, we start by attending the features provided by the mmWave subnet. Given an extracted feature vector $\mathbf{z}_M$ by the CNN extractor introduced in §5.2), a self-attention module first computes the similarity between two embedding spaces and uses the similarity to generate an attention map:

$$\mathbf{a}_M = \sigma[(\mathbf{W}_M^\varrho \mathbf{z}_M)^T \mathbf{W}_M^\varphi \mathbf{z}_M] \qquad (4)$$

where $\mathbf{W}_M^\varrho$ and $\mathbf{W}_M^\varphi$ are learnable weight matrices that project the original features to embedding spaces $\varrho(\mathbf{z}_M)$ and $\varphi(\mathbf{z}_M)$ respectively. σ represents a non-linear activation function, e.g., softmax or sigmoid. After applying the generated focus mask, the attended features $\tilde{\mathbf{z}}_M$ are given by:

$$\tilde{\mathbf{z}}_M = \mathbf{a}_M \odot \mathbf{z}_M \qquad (5)$$

where $\odot$ denotes element-wise multiplication. From the perspective of filtering, self-attention forces the model to focus on stable and geometrically meaningful features, whilst ignoring distracting or noisy components, so as to regulate the estimation model in response to complex environment uncertainty. In this way, we can similarly generate self-regulated deep odometry features for inertial sensors ($\tilde{\mathbf{z}}_I$), RGB cameras ($\tilde{\mathbf{z}}_V$) or depth cameras ($\tilde{\mathbf{z}}_D$) depending on their availability. As introduced in [61, 65], through the similarity comparison between two embeddings, self-attention is able to capture the long-range dependencies and global correlations of the features. This property is of paramount importance for mmWave and optical odometry estimation, where widely separated spatial regions should jointly considered.

6.4 Inter-sensor Cross-Attention

The self-attention layer reweighs the importance of per-sensor descriptors in the deep feature space. But such importance masks ignore the correlation among different sensors, motivating us to move forward to inter-sensor correlations. The cross-attention serves as the second-stage of reweighting, which is inspired by
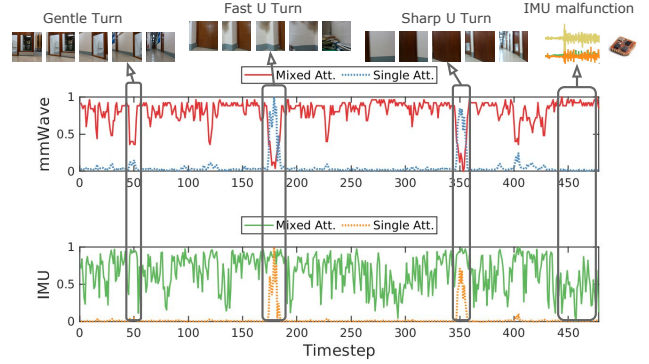

Figure 5: Attention masks comparison between Mixed (Two-stage) and Single-stage strategy. We aggregate a mask vector by counting the ratio of its sigmoid outputs over 0.5, and visualize the aggregated values over the sequence. Both attention weights are normalized to $[0, 1]$. It can be seen that with our proposed mixed strategy, mmWave and IMU *interchangeably tones down* in response to different challenging situations to them. While with the baseline single-stage method, mmWave and IMU importance change *synchronously* implying an ineffective complementary fusion.

the mechanism of human perception. Analogously, we expect a robust multi-modal egomotion system to adaptively shift relative sensor influence depending on the particular scenario in question.

To give these egomotion sensors the ability to "crosstalk", we consider generating attention masks across sensor descriptors, where the conditioning process is akin to self-attention. Consider the case combining a mmWave radar with an inertial sensor. Given their features $\tilde{\mathbf{z}}_M$ and $\tilde{\mathbf{z}}_I$ after self-attention, two respective cross-attention masks are derived as follows:

$$\mathbf{a}_{I\to M} = \sigma[(\mathbf{W}_{I\to M}^\varrho \tilde{\mathbf{z}}_I)^T \mathbf{W}_{I\to M}^\varphi \tilde{\mathbf{z}}_I]$$
$$\mathbf{a}_{M\to I} = \sigma[(\mathbf{W}_{M\to I}^\varrho \tilde{\mathbf{z}}_M)^T \mathbf{W}_{M\to I}^\varphi \tilde{\mathbf{z}}_M] \qquad (6)$$

where the family of $\mathbf{W}$ are learnable weight matrices. The final multi-modal egomotion descriptors can be modeled by concatenating the second-stage attended features:

$$\bar{\mathbf{z}}_{MI} = [\mathbf{a}_{M\to I} \odot \tilde{\mathbf{z}}_I; \mathbf{a}_{I\to M} \odot \tilde{\mathbf{z}}_M] \qquad (7)$$

When more than two egomotion sensors are available, e.g., mmWave radars ($\tilde{\mathbf{z}}_M$), inertial sensors ($\tilde{\mathbf{z}}_I$) and RGB cameras ($\tilde{\mathbf{z}}_V$), we randomly leave one out and concatenate the remainder and then generate an attention mask conditioned on it. For example, generating a mmWave attention mask from RGB and inertial sensors, can be described as:

$$\mathbf{a}_{VI\to M} = \sigma[(\mathbf{W}_{VI\to M}^\varrho \tilde{\mathbf{z}}_{VI})^T \mathbf{W}_{VI\to M}^\varphi \tilde{\mathbf{z}}_{VI}] \qquad (8)$$

where $\tilde{\mathbf{z}}_{VI} = [\tilde{\mathbf{z}}_V; \tilde{\mathbf{z}}_I]$, are the concatenated features from two sensors after self-attention. The final attended features are:

$$\bar{\mathbf{z}}_{MIV} = [\mathbf{a}_{MV\to I} \odot \tilde{\mathbf{z}}_I; \mathbf{a}_{VI\to M} \odot \tilde{\mathbf{z}}_M; \mathbf{a}_{MI\to V} \odot \tilde{\mathbf{z}}_V] \qquad (9)$$

This completes the mixed attention. The final attended features are fed into the RNN module, as discussed in §4.

6.5 Discussion of Mixed Attention

In contrast to our proposed two-stage mixed attention, a natural thought, however, is why not directly concatenate the extracted features from different odometry modules into a "big" vector, and simply perform self-attention on it (single-stage in Fig. 4). In this way, the role of the cross-attention is implicitly surrogated into a single-stage (self-)attention. We now justify why our two-stage attention is a better design choice from the perspectives of both efficacy and complexity.

Efficacy. Recall the end goal of mixed-attention is to realize adaptation through complementary sensor interactions. Although §8.2.5 will further confirm the superiority of our method based on quantitative results, a more intuitive way to examine this is by visualizing the generated masks and validating their change in behaviour across different scenarios. As we can see in Fig. 5, the masks of different sensors generated by single-stage attention do not indicate cross-modal complementariness and change *synchronously* or simultaneously over time. By inspecting the data, we find that, the concurrent weight increases are activated following the change in gyroscope readings, rather than a complementary response for adaptation. On the other hand, the focus maps generated by our two-stage attention strategy give rise to clear complementary behaviors. In particular, we can clearly see that how mmWave and IMU interchangeably dominate the estimation in response to different situations: mmWave tones down in response to small parallax due to turning, while IMU shifts roles to mmWave when encountering unexpected values caused by malfunction. We hypothesize that attention disentangling is the key to the different attention behaviors here. Again, as discussed in §6.4, the basis of complementary behavior is laid down by sensor individuality, and we achieve this by a two-stage attention strategy. Note that the self-attention stage also plays an important role in our design, which we will discuss more in §8.2.5. In a related context, disentangling self- and cross-attention is also found to be an effective manner to detect saliency from paired inputs [17].

Complexity. We further demonstrate that our two-stage attention does not increase computation cost and actually makes it simpler to train. Take the most complicated mmWave-Inertial-Visual odometry as example, where the lengths of their extracted feature vectors as N_M, N_I and N_V. Inspecting Eq. (4, 8), it is easy to find that the total weight matrix size is $2 \times (N_M + N_I + N_V)^2$ for single-stage strategy, which is same as the space complexity of our mixed attention module:

$$\underbrace{2 \times (N_M^2 + N_I^2 + N_V^2)}_{\text{1. self-attention}} + \underbrace{4 \times (N_M N_I + N_V N_I + N_M N_V)}_{\text{2. cross-attention}}$$

In fact, it has been found that by breaking learning a fat weight matrix into several smaller matrices lined up, more efficient model inference and less network over-fitting can be also obtained [58]. The runtime experimental results in §8.4 also supports this claim.

7 IMPLEMENTATION

For the purpose of reproducing our approach, we release our dataset, containing more than 8km of trajectories, and the source code of our system[3].
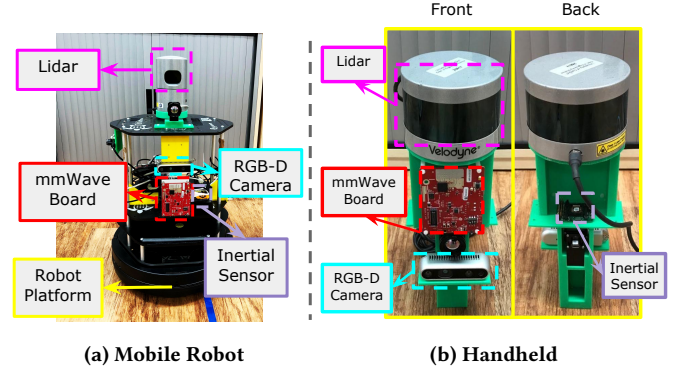
---

[3]https://github.com/ChristopherLu/milliEgo

(a) Mobile Robot   **(b) Handheld**

Figure 6: Two of the prototypes. Note the bulky lidar for providing accurate ground truth.

Multi-modal Sensing Platform. We implement our multi-modal sensing system with both a mobile robot prototype and a handheld prototype. Fig. 6a shows the robot prototype, on which we equipped a Turtlebot 2 with multiple sensors including (1) a TI AWR1843 board to collect mmWave data, (2) an Xsens MTi-1 IMU for inertial measurements, (3) an Intel D435i Depth camera for both RGB and depth image capture, and (4) a Velodyne HDL-32E lidar for ground truth labeling. All the sensors are coaxially located on the robot along the vertical axis and synchronized through ROS on Turtlebot 2 [49]. The handheld prototype is illustrated in Fig. 6b. We designed a 3D printed model that uses the same set of sensors as with the robot prototype. The only difference is the replacement of the Velodyne HDL-32E lidar with a more lightweight Velodyne Ultra Puck.

Testbeds. Our testbeds have been explicitly chosen for their wide diversity, as a common concern of deep learning techniques is their potential for overfitting. The dataset consists of 17 distinct floors from 6 different multistorey buildings, including almost all accessible areas in a commercial building, including hallways, canteen, common room, building junction, atrium, office and cluttered store rooms. The smallest floor has an area of $\sim 205 m^2$ while the largest one has an area of more than $1500 m^2$. These buildings have different internal designs and constructions and thus provide different challenges for odometry estimation, ranging from material to scale. More site descriptions can be found in our released dataset. Notably, all the experimental buildings are installed with motion-triggered energy-saving lights, which causes nontrivial dynamic illumination challenges for RGB cameras. Meanwhile, glass balustrades can be found in almost all of the buildings, which is a known issue for depth cameras [5]. We made such a testbed choice with the goal of investigating robust odometry and exploring if mmWave radar can be a cheap alternative for scenarios which are particularly challenging for optical sensors.

Data Collection Procedure. We ensure that we capture a wide diversity of trajectories, ranging from simple straight-line routes to multiple traverses of complex routes. The sampling rate of our mmWave radar is set to 20Hz based on our empirically optimal SNR configuration on the TI AWR1843 TI board. Our final dataset contains data from the mmWave radar, IMU, RGB cameras, depth cameras and lidar. Sec. 8.1 introduces how collected data are split for training and testing.

Network Training Details. For model training, we use RMSProp optimizer with a $1e-05$ initial learning rate, dropping by 25% every 25 epochs for a total of 200 epochs. We normalize the input data by subtracting the mean over the dataset. The training sequence is randomly cut into small batches of consecutive pairs ($n = 16$) to obtain better generalization. We also sub-sample the input frames to provide sufficient parallax between consecutive frames. We discuss the importance of this sampling strategy in §8.3.2. The regularization hyper-parameter γ in Eq. (2) is set to 0.001. The entire training time of `milliEgo` is around 6 hours on a NVIDIA Tesla V100.

8 EVALUATION

8.1 Experimental Setting

Metrics and Ground Truth. To evaluate the proposed model, we follow a widely used odometry benchmark [23] and adopt the Absolute Trajectory Error (*ATE*) to quantify the tracking accuracy of an entire trajectory. The reason we use the entire trajectory is because the egomotion between consecutive frames is often too small to be numerically significant for the drifting effect. In contrast, ATE is considered a holistic metric that measures the composed long-term odometry performance. In particular, we report multiple statistics of ATE, including mean Root Mean Square Error (Mean), standard deviation (Std.) and max error (Max). To obtain the ground truth, we use a co-located lidar on the mobile platform and run laser-based SLAM (i.e., gmapping [53] in our case) to get its 6-DoF pose as our ground truth, following the practice of [54].

Competing Approaches. The baselines include both conventional and deep learning based methods. For the baseline combination of mmWave+IMU, we compare with the *IMU-ICP* [1], which is a **traditional** fusion method that uses inertial sensors to bootstrap the ICP registration of the mmWave data via the adaptive Kalman filter. We also compare with other state-of-the-art multi-modal deep odometry with different sensors, including *VINET* for RGB camera plus IMU and its variant *DINET* for depth camera plus IMU. On top of these multi-modal baselines, we also compare with RANSAC-IMU [14] and TEASER [72] in §8.2.3 for examining standalone mmWave radar odometry i.e. without inertial aiding.

Evaluation Protocol. Both datasets of mobile robot and handheld device are divided into training and test sets. 5-fold cross-validation is used to select the best model for testing. A key principle we follow in dividing them is to ensure that there is sufficient diversity between test and training sets to demonstrate generalization i.e. with distinct motion traces or tests on unseen environments. In this way, the model generalization ability can be fairly examined. Concretely, for the robot-platform evaluation, we train the model with 49 training sequences and test the results on 6 held-out sequences. We train another model with 27 sequences for the handheld-device evaluation and test it on 3 held-out sequences. The dataset is substantial in size: for the robot, the training set is 2878 m and the test set is 438 m. For the handheld case, the training set is 4380 m and the test set is 315 m.

8.2 Mobile Robot Performance

We start by evaluating the odometry performance on mobile robot as our primary platform.

Table 1: Overall results on the mobile robot. (Unit - Meter)

Sensors	IMU Only	mmWave + IMU		RGB + IMU	Depth + IMU
Method	IONET	IMU-ICP	`milliEgo`	VINET	DINET
3D Mean	2.644	5.054	**0.814**	1.955	2.255
Std.	1.843	1.742	**0.444**	1.479	1.526
Max	6.402	7.866	**1.689**	5.633	5.53
2D Mean	2.45	5.008	**0.764**	1.936	2.239
Std.	1.748	1.745	**0.438**	1.478	1.525
Max	6.011	7.807	**1.629**	5.608	5.514
Params (M)	1.5	-	33.9	190.9	190.9

Table 2: Extending `milliEgo` to triple-sensor fusion. Sensor Key: *M* - mmWave, *I* - inertial, *V* - RGB, *D* - Depth.

Method	3D			2D		
	Mean	**Std.**	**Max**	**Mean**	**Std.**	**Max**
M + I + V (w.o. Att.)	0.862	0.469	1.871	0.838	0.458	1.838
M + I + D (w.o. Att.)	1.194	0.791	2.791	1.008	0.746	2.563
M + I + V	0.702	0.399	1.608	0.673	0.381	1.552
M + I + D	0.769	0.498	1.821	0.761	0.496	1.813

8.2.1 Overall Performance. Tab. 1 summarizes the overall performance in 2D and 3D space with fusion of two sensor modalities. One of the sensors is fixed as an IMU due its pervasiveness on modern mobile platforms. We compare against RGB+IMU and Depth+IMU, using benchmark deep learning approaches. For comparison, the odometry performance using only an IMU alone is also shown.

As can be seen, `milliEgo` achieves the lowest error amongst all methods, yielding a 3D ATE of 0.814m, approximately a 1.3% error in drift when taking the trajectory distances into account. This substantially outperforms the IMU-ICP methods by almost an order of magnitude, which shows the worst performance, even worse than IMU alone. This is due to its inability to deal with the low-quality mmWave point clouds. `milliEgo` also largely outperforms the other dual-sensor systems like RGB+IMU and depth+IMU. As discussed in the §7, our datatset contains challenging illumination conditions that occur in the real world (e.g., glare, dimness, darkness and glass windows) that impede their performance. Conversely, as mmWave is insensitive to ambient illumination, it is able to provide more accurate odometry estimation. `milliEgo`'s ATE is further reduced to 0.764m in the 2D plane, which allows for sub-metre indoor tracking.

8.2.2 Extending to Triple Sensor Egomotion. `milliEgo` can also be extended to triple-sensor egomotion systems. On the basis of §8.2.1, we use mmWave and IMU as the two baseline sensors and choose either the RGB camera or depth camera as the third, as shown in Tab. 2. When comparing to Tab. 1, an interesting tradeoff between system complexity and performance can be noticed. Although extending `milliEgo` to triple-sensor egomotion improves the accuracy, the delta is not substantial. Compared to the two-sensor version, only $\sim$ 13% and $\sim$ 6% ATE reduction are observed after incorporating an extra RGB and depth camera respectively. This, however, is at the
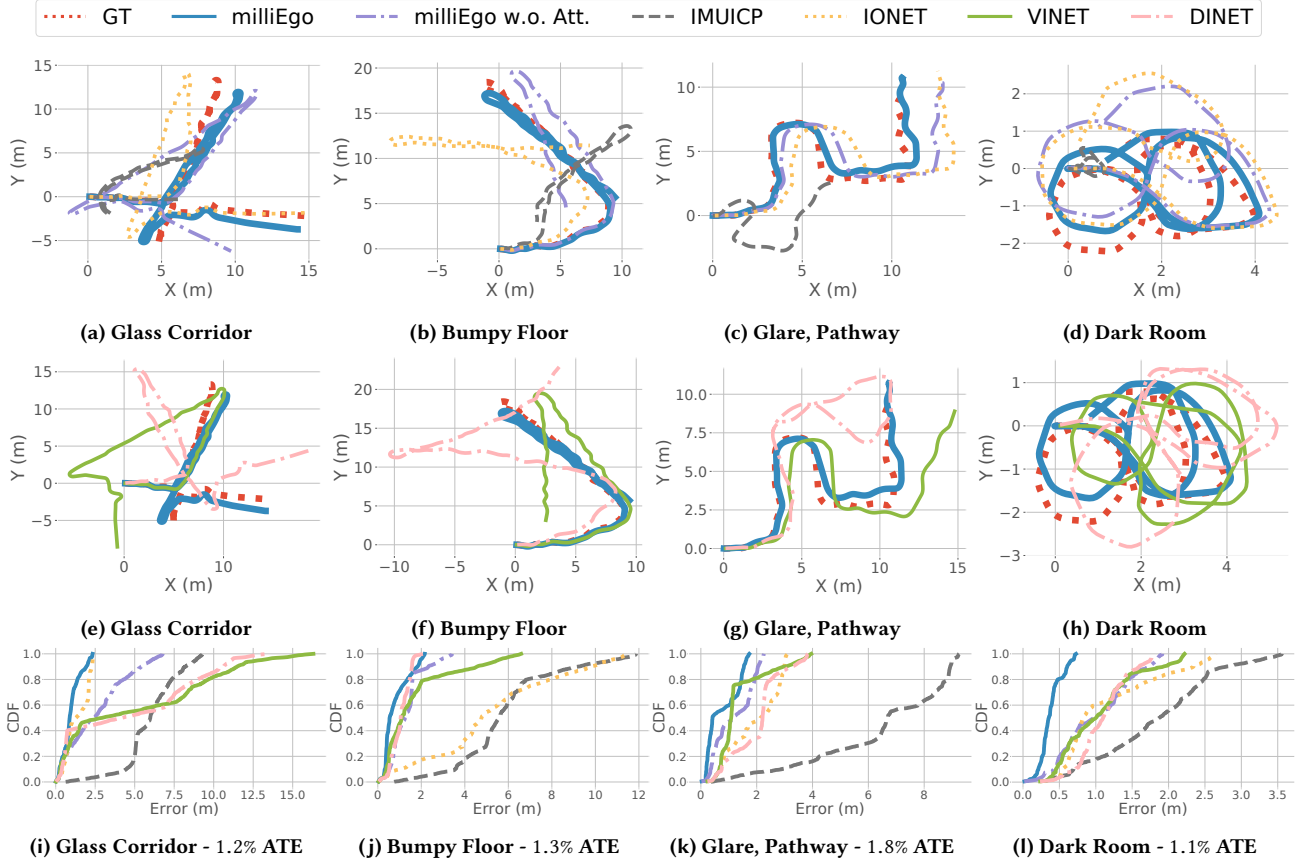
Figure 7: Prediction comparison on four exemplar testing trajectories with the robot platforms. Top row: method-level comparison; Middle row: sensor-level comparison; Bottom row: cumulative distribution function (CDF) plots.

cost of hardware/sensor overhead and computational latency. We thus suggest that end users carefully consider this tradeoff before extending to three sensors.

8.2.3 Effectiveness of mmWave Subnet. We next examine the effectiveness of the mmWave subnet alone, i.e. without the use of the IMU or attention, as discussed in Sec. 5.2. We compare the performance of standalone subnet with two **conventional** point registration methods: (1) *ICP* and (2) *TEASER*. ICP [14] is the widely adopted point registration method to estimate egomotion. We follow the conventional practice that uses RANSAC algorithm [30] for outlier rejection. TEASER [72] is the latest state-of-the-art method using a sophisticated optimization to achieve robust point registration against strong outliers. Due to the very sparse and noisy mmWave point clouds, these two baselines unfortunately cannot succeed for most testing sequences (refer to §5.1). We therefore consider only 3 out of our 7 testing sequences on which the baselines can give reasonable odometry estimation. As reference, we also consider inertial only odometry (IONET).

Results. As can be seen in Fig. 8, the mmWave subnet surpasses both baselines by at least 3-fold on all testing sequences. In particular, we found that the average 2D and 3D ATE are only 2.47m and
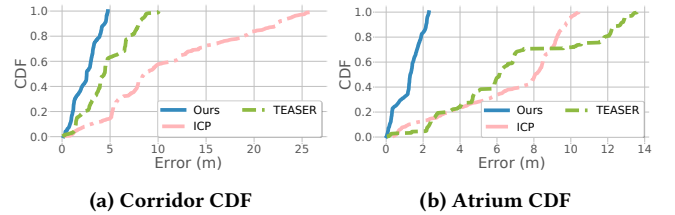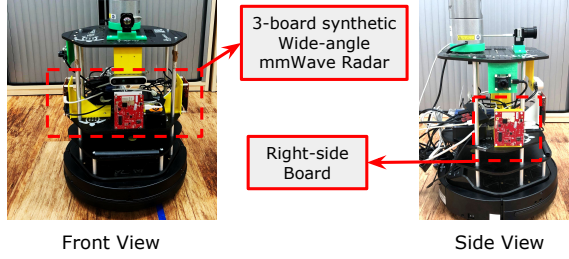

Figure 8: CDF of two test sequences with mmWave Subnet.

2.50m. In contrast, ICP and TEASER struggle in providing reliable egomotion estimation due to the sparse and noisy point clouds.

8.2.4 Potential impact of multiple mmWave radars. As an emerging technology, the Field of View (FoV) of current single-chip mmWave radars is still limited. However, given the promise of such technology and the current practice of installing multiple mmWave radars on an automobile [32], it is worth studying the potential impact of multiple co-located radars that collectively form a wide-angle sensing view. To fully understanding this impact, we evaluate the *mmWave-only* version of `milliEgo` on 5 out of the 7 test sequences, in which there were 3 radars co-located on the robot (front, left

Table 3: Investigation into different attention strategies

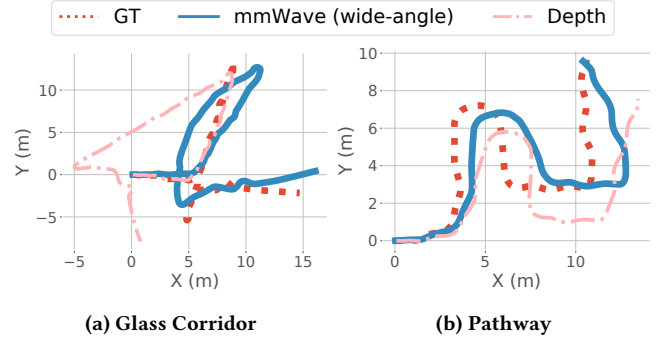| Method		No Att.	Single-Stage Att. [13, 54]	w.o. Cross Att.	w.o. Self Att.	milliEgo
3 D	Mean	1.373	1.494	0.923	0.949	**0.814**
	Std.	0.784	0.838	0.437	0.532	**0.444**
	Max	2.981	3.122	1.893	1.971	**1.689**
2 D	Mean	1.369	1.441	0.91	0.935	**0.764**
	Std.	0.781	0.793	**0.437**	0.528	0.438
	Max	2.976	3.027	1.883	1.958	**1.629**
Params (M)		31.6	42.7	32.3	33.3	33.9

Front View Side View

Figure 9: The wide-angle mmWave radar synthesized by 3 co-located radars.

Table 4: Performance of a 'wide-angle' mmWave radar system.

Sensors		mmWave		Depth	IMU	RGB
		Narrow	Wide Angle			
3 D	Mean	2.611	**1.239**	1.882	2.285	1.983
	Std.	1.396	**0.783**	1.220	1.134	1.312
	Max	4.641	**2.990**	4.580	4.390	4.673
2 D	Mean	2.507	**1.208**	1.879	2.275	1.973
	Std.	1.218	**0.792**	1.220	1.137	1.309
	Max	4.210	**2.979**	4.578	4.384	4.658

and right sides as shown in Fig. 9). This collocation changes the horizontal FoV from 120 deg to 240 deg. To align the radars, we determine the transformation to the center of our mobile robot, and then stitch the three transformed point clouds into a single, merged cloud. For baselines, we compare its egomotion estimation against other sensor modalities: RGB, depth and inertial.

Results. As can be seen in Tab. 4, the 'wide-angle' view created by co-located radars significantly reduces the ATE of mmWave-only odometry over the single-radar case (i.e., normal view). Meanwhile, wide-angle mmWave-only odometry outperforms all other sensor modalities on every dimension of the ATE metric. Compared to the best baseline (i.e., Depth camera only), wide-angle milliEgo reduces mean ATE by 35%. Fig. 10 shows two exemplar trajectory comparisons between wide-angle mmWave and depth camera. This experiment demonstrates that a wide-angle view (e.g. through increased number of MIMO antenna elements) will lead to improved performance. Another takeaway from Tab. 4 is that all single-modal odometry systems have their limitations in egomotion estimation, reinforcing the need for a multi-modal solution.

8.2.5 Effectiveness of Mixed Attention. Lastly we quantitatively validate our mixed attention introduced in §6.2 for multi-modal

(a) Glass Corridor (b) Pathway

Figure 10: Two example testing sequences with wide-angle mmWave odometry alone. For clarity, we only show the trajectory for depth camera as the most competitive alternative in Tab. 4.

sensor fusion. Specifically, we investigate different fusion strategies in milliEgo, including (1) No attention, (2) single-stage attention (see §6.5), (3) milliEgo without self-attention (4) milliEgo without cross-attention. Tab. 3 suggests that our mixed attention strategy brings ∼ 40% performance gain compared to milliEgo without attention. Moreover, it consistently outperforms other attention strategies in both 2D and 3D planes, providing more than a 50% accuracy increase over single-stage attention. This is consistent with the analysis in §6.5. Moreover, the single-stage attention model has more parameters than milliEgo due to its 'fat' mask generation on the concatenated vector. On the other hand, it can be noticed from Tab. 3 that every stage of attention contributes; removing either one causes sub-optimal performance. Interestingly, the ATE performances of self-attention only and cross-attention only are very close, suggesting their equal importance.

8.3 Handheld Device Performance

Ground robots move in a horizontal plane and therefore are relatively constrained for egomotion estimation. We therefore investigate the performance of milliEgo in a less constrained case i.e. carried in the hand.

8.3.1 Overall Performance. Tab. 5 summarizes overall performance. As can be seen, milliEgo clearly surpasses the other methods in both 3D and 2D planes. It yields an average 3D ATE of 1.895m, equivalent to an 1.8% trajectory drift. In the 2D space, its error is further reduced to 1.252m.

As shown in Fig. 11, our method copes well with different levels of trajectory complexity and constantly provides accurate odometry estimation. This confirms the versatility of our proposed framework to different mobility patterns and constraints.

8.3.2 Impact of mmWave Sampling Rate. The key to accurate egomotion estimation lies in the extent of *parallax/baseline* which is the distance between successive frames when moving [55]. Intuitively, when a fast-moving platform is equipped with a low frame-per-second (FPS) sensor, the parallax is large, resulting in less information correspondence, and ultimately a large error. Due to the low-quality mmWave point cloud, an adequate sampling rate is vital. Fig. 12 illustrates the ATE of milliEgo trained with different

Table 5: Overall results with a handheld device.

Sensors		IMU Only	mmWave + IMU		RGB + IMU
Method		IONET	IMU-ICP	milliEgo	VINET
3 D	Mean	3.452	5.843	**1.857**	4.044
	Std.	1.769	2.966	**1.026**	1.847
	Max	7.071	11.712	**4.193**	7.734
2 D	Mean	3.406	4.979	**1.459**	3.741
	Std.	1.717	2.78	**0.815**	1.854
	Max	7.055	10.928	**3.990**	7.424

(a) Corridor Trajectory

(b) Atrium Trajectory
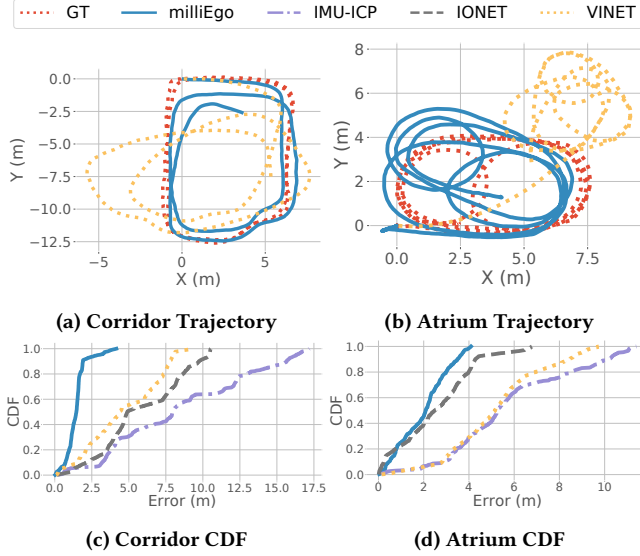
(c) Corridor CDF

(d) Atrium CDF

Figure 11: Two example testing sequences with handheld data. For clarity, we only show the trajectory for VINET in the top-row as the most competitive alternative.
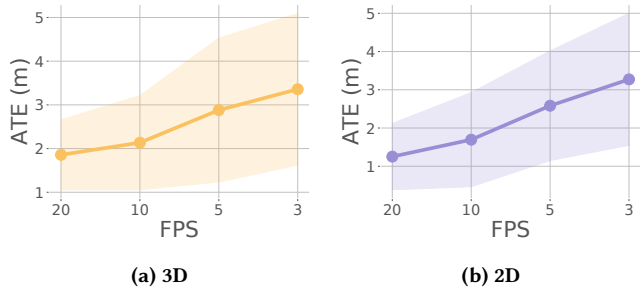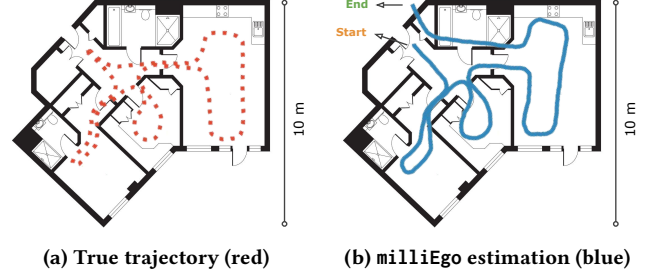

(a) 3D

(b) 2D

Figure 12: Impact of sampling rate with handheld data.

mmWave sampling rates. As can be seen, our model gives the best performance with a 20 FPS rate, with consistent performance decline with lower rates, indicating the importance of an adequate sampling rate. Notably, the 20 FPS adopted here is 5-fold as large as the robot's 4 FPS, which is reasonable due to difference in speed between the robot and a human walking. In practice, we suggest end users set the the best FPS based on the expected platform dynamics.

(a) True trajectory (red)

(b) milliEgo estimation (blue)

Figure 13: milliEgo examined in an *unseen* household environment operating in hand-held mode. Note that the device was carried by a taller user *not* involved in the original training data collection.

8.3.3 Generalization to Unseen Environments. We further evaluate the generalization ability of milliEgo in unseen scenarios. Although the handheld milliEgo model was trained on data collected from commercial buildings, we examine its egomotion estimation performance in an unseen household environment, where the layout and floor plan are significantly different to the training case. To fully validate the generalization, a new user (height of 185cm) carried the handheld device rather than the two shorter users in the training set (166cm and 178cm respectively). Fig. 13 shows the trajectory estimated by milliEgo. Despite the previously unseen environment, milliEgo demonstrates an impressive generalization performance: the mean ATE is only 0.588m with the largest ATE being only 1.964m. Note that in contrast deep visual odometry in particular can be poor at adapting to new environments with different visual features. By inspecting Fig. 13, we found that the main trajectory is well revealed, implying a very good precursor for downstream tasks such as SLAM or map-matching based localization [68].

8.4 System Efficiency

In the last experiment, we implement and compare milliEgo on two platforms: an NVIDIA Jetson TX2 (TX2) and a mini netbook. For the implementation, we use TensorFlow Lite [4] to compress our models as per the convention of efficient on-device inference of DNNs. We focus on the comparison with VINET, which is a pervasive DNN solution for estimating egomotion on resource-constrained platforms. Tab. 6 suggests that our proposed milliEgo is significantly faster (17 FPS on TX2) than VINET (3 FPS on TX2). This is because the transformed mmWave panoramic images are much smaller and hence more data-efficient than RGB images. Note also that compared with the single-stage attention strategy, milliEgo is slightly faster as it divides the single, large mask generation process into several lightweight branches (see §6.5). Concerning the model size (in .tflite format), milliEgo is only 107MB as opposed to the 171MB for single-stage attention and 764MB of VINET. This further validates milliEgo's efficiency and suitability for more resource constrained platforms.

9 RELATED WORK

Radar based Localization and Tracking. Prior work in this thread can be broadly divided into three categories based on the class of

Table 6: Runtime Analysis (Unit: Second).

	milliEgo	VINET
TX2	0.0588 ± 0.0028	0.2392 ± 0.0027
Notebook	0.484 ± 0.009	10.398 ± 0.035

radar used: (1) surveillance radar, (2) mechanical scanning radar and (3) single-chip mmWave radar. Multistatic or bistatic surveillance radar is one of the most well-known technologies in the radar community and has been widely used in military for aircraft tracking [7, 42, 48]. These are long range, high-power with custom designed signal processing chains. In the second class, to meet the mobility needs of automobiles, mechanical scanning radars (e.g., Navtech CTS350-X) are used for odometry estimation [9, 46]. However, as discussed in [38], mechanical radars are overly heavy and cumbersome for service robots or wearable devices. Single-chip mmWave radars have recently emerge as a low-cost and small-factor solution to the above radars and have been used in indoor mapping [38] and ego-velocity estimation [31]. Compared with prior art, our work uses the same single-chip radar but focuses on different tasks. As a result, milliEgo has its own technical challenge in a new context and leverages a multi-sensor solution to robustly estimate the trajectory of a mobile platform.

Deep Odometry Model. Owning to recent advances in data-driven methods, deep odometry models have gained attention in both academia and industry. A multitude of such models have been proposed and are widely used with different sensors, including RGB camera [35, 62], depth camera [59, 73], IMU [11, 12] and lidar [44, 64]. A prominent feature of deep odometry lies in its end-to-end learning, which enables the modeling of many complex physical processes and disturbances [63]. Beside this benefit, using deep odometry models can enable the system to skip the tedious calibration phase, increasing the adoption range even for non-domain-experts. Our work also leverage the superior modeling power of deep odometry model. Compared with the above works, we develop the first deep mmWave odometry milliEgo with the low-cost single chip radar, and systematically examined its limitation and potential to assist other odometry sensors.

Multi-modal Sensor Fusion Our work also falls within the scope of multi-modal sensor fusion, a technique that combines the information from multiple sensor modalities in order to provide better system performance. Different to traditional sensor fusion methods (e.g., probabilistic fusion, fuzzy reasoning, hybrid fusion, etc [3, 10, 69]), neural network based sensor fusion methods are able to fuse heterogeneous sensor data at all levels ranging from raw data, to intermediate features, to individual decisions. This uniqueness makes deep sensor fusion an emerging topic in recent years. Early works [20, 74] in this vein treat various sensor modalities equally in deep neural networks yet ignore their heterogeneous sensing qualities. By incorporating self-attention mechanisms [61] or similar concept [65] in the network, recent efforts [36, 70, 75?] made progress in dynamically understanding the importance of inputs or features and robustly recognize human activity. In the domain of egomotion estimation, such a single-stage attention mechanism has also been used in multi-modal odometry to adaptively combine inertial sensors with RGB or thermal imaging cameras (e.g.,

SelectiveFusion [13] and DeepTIO [54]). milliEgo differs from the above work in several aspects. First, milliEgo is a two-stage attention fusion. It not only uses a self-attention layer to prioritize intra-sensor features, but further proposes a cross-attention layer based on the Spence's spatial perception model [57] that enforces complementary inter-sensor behaviors towards achieving greater robustness. As shown by the experiment results, milliEgo significantly outperforms [13, 54] for egomotion estimation and yield more explainable behaviors between mmWave radars and other sensors. Secondly, milliEgo is the first work that end-to-end fuses a low-cost mmWave radar with other sensors in a neural network.

10 LIMITATIONS AND FUTURE WORK

This work focuses on a proof-of-principle egomotion estimation assisted by single-chip mmWave radar, towards our vision of enabling a high degree of spatial awareness with low-cost mobile sensing systems. There are limitations and a number of avenues for future exploration. Firstly, as our design and experiments have mainly focused on indoor commercial buildings, it remains unknown how milliEgo will generalize to other indoor spaces (e.g, hospitals, museums and shopping malls) or even outdoor scenarios. This will likely require more diverse data collection for account for the distinct noise in different environments. The optimal range settings will also need to be investigated as objects are likely to be more widely dispersed. Secondly, trials on UAVs or drone platforms are needed to investigate if their rapid flying speed, long parallax and less constrained motion will cause new estimation challenges. Thirdly, as with any kind of odometry, milliEgo still drifts. Therefore an obvious next step is to detect and use mmWave loop closures to refine milliEgo, so that a robust system for simultaneous localization and mapping can be achieved. In addition, although we have investigated an end-to-end system e.g. using a Jetson TX2 embedded platform for real-time odometry, there is significant room for further optimization and testing the energy efficiency on mobile devices. Lastly, it would be interesting to explore how current and emerging 60GHz MIMO WiFi standards e.g. 802.11ay/802.11az [77] can be adapted to give similar spatial information. More powerful MIMO mmWave chipsets with higher spatial and velocity resolution will also be investigated for tracking and egomotion estimation.

11 CONCLUSION

This paper presents milliEgo, a novel low-cost mmWave radar assisted odometry system that can robustly estimate egomotion together with an IMU. To address the limitation brought by the low-cost radar, milliEgo proposed to use a mmWave odometry subnet to directly learn the motion transformation from noisy data. milliEgo also features a cross-attention based multi-modal fusion that effectively combines mmWave radar with inertial and other potentially available sensors. Extensive experiments were conducted with real-time prototype implementation on both robots and handheld devices. milliEgo opens up mmWave-assisted sensing to new applications where egomotion is crucial, such as robot tracking, VR/AR/MR gaming and smartphone indoor navigation.

REFERENCES

[1] Farhad Aghili and Chun-Yi Su. 2016. Robust relative navigation by integration of ICP and adaptive Kalman filter using laser scanner and IMU. *IEEE/ASME Transactions on Mechatronics* 21, 4 (2016), 2015–2026.

[2] Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. 2016. Learning to poke by poking: Experiential learning of intuitive physics. In *Advances in neural information processing systems*.

[3] Yasin Almalioglu, Mehmet Turan, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham. 2020. Milli-RIO: Ego-Motion Estimation with Low-Cost Millimetre-Wave Radar. *IEEE Sensors Journal* (2020).

[4] Oscar Alsing. 2018. Mobile Object Detection using TensorFlow Lite and Transfer Learning.

[5] Nicolas Alt, Patrick Rives, and Eckehard Steinbach. 2013. Reconstruction of transparent objects in unstructured scenes with a depth camera. In *2013 IEEE International Conference on Image Processing*. 4131–4135.

[6] Sofien Bouaziz, Andrea Tagliasacchi, and Mark Pauly. 2013. Sparse iterative closest point. In *Computer graphics forum*, Vol. 32. 113–123.

[7] Carlos Pena Caballero, Elifaleth Cantu, Jesus Rodriguez, Adolfo Gonzales, Osvaldo Castellanos, Angel Cantu, Megan Strait, Jae Son, and Dongchul Kim. [n. d.]. A Multiple Radar Approach for Automatic Target Recognition of Aircraft using Inverse Synthetic Aperture Radar. In *2018 1st International Conference on Data Intelligence and Security (ICDIS)*.

[8] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. 2016. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics* (2016).

[9] Sarah H Cen and Paul Newman. 2018. Precise ego-motion estimation with millimeter-wave radar under diverse and challenging conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*.

[10] Balasubramaniyan Chandrasekaran, Shruti Gangadhar, and James M Conrad. 2017. A survey of multisensor fusion techniques, architectures and methodologies. In *SoutheastCon 2017*. IEEE, 1–8.

[11] Changhao Chen, Xiaoxuan Lu, Andrew Markham, and Niki Trigoni. 2018. Ionet: Learning to cure the curse of drift in inertial odometry. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[12] Changhao Chen, Xiaoxuan Lu, Johan Wahlstrom, Andrew Markham, and Niki Trigoni. 2019. Deep Neural Network Based Inertial Odometry Using Low-cost Inertial Measurement Units. *IEEE Transactions on Mobile Computing* (2019).

[13] Changhao Chen, Stefano Rosa, Yishu Miao, Chris Xiaoxuan Lu, Wei Wu, Andrew Markham, and Niki Trigoni. 2019. Selective sensor fusion for neural visual-inertial odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[14] Javier Civera, Oscar G Grasa, Andrew J Davison, and JMM Montiel. 2010. 1-Point RANSAC for extended Kalman filtering: Application to real-time structure from motion and visual odometry. *Journal of field robotics* 27, 5 (2010), 609–631.

[15] Ronald Clark, Sen Wang, Hongkai Wen, Andrew Markham, and Niki Trigoni. 2017. Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[16] George E Dahl, Tara N Sainath, and Geoffrey E Hinton. 2013. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *2013 IEEE international conference on acoustics, speech and signal processing*. 8609–8613.

[17] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 224–236.

[18] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. 2015. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 2758–2766.

[19] Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Longyue Wang, Shuming Shi, and Tong Zhang. 2019. Dynamic layer aggregation for neural machine translation with routing-by-agreement. In *AAAI*.

[20] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. 2015. Multimodal deep learning for robust RGB-D object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 681–687.

[21] Steven Gold, Anand Rangarajan, Chien-Ping Lu, Suguna Pappu, and Eric Mjolsness. 1998. New algorithms for 2D and 3D point matching: Pose estimation and correspondence. *Pattern recognition* 31, 8 (1998), 1019–1031.

[22] Davide Guermandi, Qixian Shi, Andy Dewilde, Veerle Derudder, Ubaid Ahmad, Annachiara Spagnolo, Ilja Ocket, André Bourdoux, Piet Wambacq, Jan Craninckx, et al. 2017. A 79-GHz MIMO PMCW radar SoC in 28-nm CMOS. *IEEE Journal of Solid-State Circuits* 52, 10 (2017), 2613–2626.

[23] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. 2014. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *IEEE international conference on Robotics and automation (ICRA)*.

[24] Gibson Hu, Shoudong Huang, Liang Zhao, Alen Alempijevic, and Gamini Dissanayake. 2012. A robust rgb-d slam algorithm. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 1714–1719.

[25] Daniel P Huttenlocher and Shimon Ullman. 1990. Recognizing solid objects by alignment with an image. *International journal of computer vision* 5, 2 (1990), 195–212.

[26] Texas Instruments. [n. d.]. Automotive mmWave sensors. http://www.ti.com/sensors/mmwave/overview.html

[27] Texas Instruments. [n. d.]. mmWave sensors Overview. https://www.ti.com/sensors/mmwave-radar/overview.html

[28] Texas Instruments. [n. d.]. mmWave Training Series. https://training.ti.com/mmwave-training-series

[29] Jean-Pierre Joosting. [n. d.]. Radar enables heavy lifting drones to navigate in complex environments. https://www.eenewseurope.com/news/radar-enables-heavy-lifting-drones-navigate-complex-environments?from=singlemessage&isappinstalled=0#

[30] Deok-Hwa Kim and Jong-Hwan Kim. 2013. Image-Based ICP algorithm for visual odometry using a RGB-D sensor in a dynamic environment. In *Robot Intelligence Technology and Applications 2012*. 423–430.

[31] Andrew Kramer, Carl Stahoviak, Angel Santamaria-Navarro, Ali-akbar Agha-mohammadi, and Christoffer Heckman. 2020. Radar-Inertial Ego-Velocity Estimation for Visually Degraded Environments. In *IEEE International Conference on Robotics and Automation (ICRA)*.

[32] MARK LAPEDUS. 2017. Here Comes High-Res Car Radar. https://semiengineering.com/here-comes-high-res-car-radar/

[33] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. 2015. Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research* 34, 3 (2015), 314–334.

[34] Bo Li, Tianlei Zhang, and Tian Xia. 2016. Vehicle detection from 3d lidar using fully convolutional network. *arXiv preprint arXiv:1608.07916* (2016).

[35] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. 2018. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *2018 IEEE international conference on robotics and automation (ICRA)*. 7286–7291.

[36] Shengzhong Liu, Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Huajie Shao, and Tarek Abdelzaher. 2020. GIobalFusion: A Global Attentional Deep Learning Framework for Multisensor Information Fusion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–27.

[37] Chris Xiaoxuan Lu, Yang Li, Peijun Zhao, Changhao Chen, Linhai Xie, Hongkai Wen, Rui Tan, and Niki Trigoni. 2018. Simultaneous localization and mapping with power network electromagnetic field. In *Proceedings of the 24th annual international conference on mobile computing and networking (MobiCom)*.

[38] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A. Stankovic, Niki Trigoni, and Andrew Markham. 2020. See Through Smoke: Robust Indoor Mapping with Low-cost mmWave Radar. In *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*.

[39] Jonathan S Lu, Patrick Cabrol, Daniel Steinbach, and Ravikumar V Pragada. 2013. Measurement and characterization of various outdoor 60 GHz diffracted and scattered paths. In *MILCOM*.

[40] Dimitrios Lymberopoulos, Jie Liu, Xue Yang, Romit Roy Choudhury, Vlado Handziski, and Souvik Sen. 2015. A realistic evaluation and comparison of indoor location technologies: Experiences and lessons learned. In *Proceedings of the 14th international conference on information processing in sensor networks*.

[41] George R MacCartney, Sijia Deng, Shu Sun, and Theodore S Rappaport. 2016. Millimeter-wave human blockage at 73 GHz with a simple double knife-edge diffraction model and extension for directional antennas. In *IEEE Vehicular Technology Conference (VTC-Fall)*.

[42] Babak Mamandipoor, Greg Malysa, Amin Arbabian, Upamanyu Madhow, and Karam Noujeim. 2014. 60 ghz synthetic aperture radar for short-range imaging: Theory and experiments. In *ACSSC*.

[43] Andriy Myronenko and Xubo Song. 2010. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence* 32, 12 (2010), 2262–2275.

[44] Austin Nicolai, Ryan Skeele, Christopher Eriksen, and Geoffrey A Hollinger. 2016. Deep learning for laser based odometry estimation. In *RSS workshop Limits and Potentials of Deep Learning in Robotics*.

[45] Chris Nuttall and Tim Bradshaw. [n. d.]. Google draws on old radar technology for its motion sensor Pixel 4 smartphone. https://www.ft.com/content/02c051ec-f005-11e9-ad1e-4367d8281195

[46] Yeong Sang Park, Young-Sik Shin, and Ayoung Kim. 2020. PhaRaO: Direct Radar Odometry using Phase Correlation. In *IEEE ICRA*.

[47] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image transformer. In *ICML*.

[48] Michail N Petsios, Emmanouil G Alivizatos, and Nikolaos K Uzunoglu. 2008. Solving the association problem for a multistatic range-only radar target tracker. *Signal Processing* (2008).

[49] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source Robot Operating System. In *ICRA workshop on open source software*, Vol. 3. 5.

[50] Valentin Radu and Mahesh K Marina. 2013. HiMLoc: Indoor smartphone localization via activity aware pedestrian dead reckoning with selective crowdsourced

WiFi fingerprinting. In *International conference on indoor positioning and indoor navigation*.

[51] Jeffrey J Richardson and L Monika Moskal. 2011. Strengths and limitations of assessing forest density and spatial configuration with aerial LiDAR. *Remote Sensing of Environment* 115, 10 (2011), 2640–2651.

[52] Peng Rong and Mihail L Sichitiu. 2006. Angle of arrival localization for wireless sensor networks. In *SECON*.

[53] ROS.org. [n. d.]. gmapping. http://wiki.ros.org/gmapping

[54] Muhamad Risqi U Saputra, Pedro Porto Buarque de Gusmao, Chris Xiaoxuan Lu, Yasin Almalioglu, Stefano Rosa, Changhao Chen, Johan Wahlstrom, Wei Wang, Andrew Markham, and Niki Trigoni. 2020. Deeptio: A deep thermal-inertial odometry with visual hallucination. *IEEE Robotics and Automation Letters* (2020).

[55] Muhamad Risqi U Saputra, Andrew Markham, and Niki Trigoni. 2018. Visual SLAM and structure from motion in dynamic environments: A survey. *ACM Computing Surveys (CSUR)* 51, 2 (2018).

[56] Sheng Shen, Mahanth Gowda, and Romit Roy Choudhury. 2018. Closing the gaps in inertial motion tracking. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 429–444.

[57] Charles Spence and Jon Driver. 2004. *Crossmodal space and crossmodal attention.* Oxford University Press.

[58] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Advances in neural information processing systems*. 2377–2385.

[59] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. 2017. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[60] Deepak Uttam and B Culshaw. 1985. Precision time domain reflectometry in optical fiber systems using a frequency modulated continuous wave ranging technique. *Journal of Lightwave Technology* (1985).

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*.

[62] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. 2017. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 2043–2050.

[63] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. 2018. End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks. *The International Journal of Robotics Research* (2018).

[64] Wei Wang, Muhamad Risqi U Saputra, Peijun Zhao, Pedro Gusmao, Bo Yang, Changhao Chen, Andrew Markham, and Niki Trigoni. 2019. DeepPCO: End-to-End Point Cloud Odometry through Deep Parallel Neural Network. *International Conference on Intelligent Robots and Systems* (2019).

[65] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *CVPR*.

[66] DK Barton HR Ward. 1969. *Handbook of radar measurement.*

[67] Rob Weston, Sarah Cen, Paul Newman, and Ingmar Posner. 2018. Probably unknown: Deep inverse sensor modelling in radar. In *ICRA*.

[68] Zhuoling Xiao, Hongkai Wen, Andrew Markham, and Niki Trigoni. 2014. Light-weight map matching for indoor localisation using conditional random fields. In *Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*.

[69] Qiang Xu, Rong Zheng, and Steve Hranilovic. 2015. IDyLL: Indoor localization using inertial and light sensors on smartphones. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

[70] Hongfei Xue, Wenjun Jiang, Chenglin Miao, Ye Yuan, Fenglong Ma, Xin Ma, Yijiang Wang, Shuochao Yao, Wenyao Xu, Aidong Zhang, et al. 2019. DeepFusion: A Deep Learning Framework for the Fusion of Heterogeneous Sensory Data. In *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 151–160.

[71] Yan Yan, Long Li, Guodong Xie, Changjing Bao, Peicheng Liao, Hao Huang, Yongxiong Ren, Nisar Ahmed, Zhe Wang, et al. 2016. Multipath effects in millimetre-wave wireless communication using orbital angular momentum multiplexing. *Scientific reports* 6 (2016), 33482.

[72] Heng Yang, Jingnan Shi, and Luca Carlone. 2020. TEASER: Fast and Certifiable Point Cloud Registration. *arXiv preprint arXiv:2001.07715* (2020).

[73] Nan Yang, Rui Wang, Jorg Stuckler, and Daniel Cremers. 2018. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 817–833.

[74] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th International Conference on World Wide Web*.

[75] Shuochao Yao, Yiran Zhao, Huajie Shao, Dongxin Liu, Shengzhong Liu, Yifan Hao, Ailing Piao, Shaohan Hu, Lu Su, and Tarek F Abdelzaher. 2019. SADeepSense: Self-Attention Deep Learning Framework for Heterogeneous On-Device Sensors in Internet of Things Applications. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*.

[76] Xiaoqing Ye, Jiamao Li, Hexiao Huang, Liang Du, and Xiaolin Zhang. 2018. 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[77] Renjie Zhao, Timothy Woodford, Teng Wei, Qian Kun, and Xinyu Zhang. 2020. M-Cube: A Millimeter-Wave Massive MIMO Software Radio. In *ACM MobiCom*.