A Picture is Worth 1,000 Millimeters: Combining Vision and Wi-Fi to Improve Localization

Shazal Irshad, Eric Rozner
University of Colorado Boulder
shazal.irshad, eric.rozner@colorado.edu

Apurv Bhartia, Bo Chen
Cisco Meraki
apurv, bochen2@cisco.com

***Abstract*—In the past, researchers designed, deployed, and evaluated Wi-Fi based localization techniques in order to locate users and devices without adding extra or costly infrastructure. However, as infrastructure deployments change, one must re-examine the role of Wi-Fi localization. Today, cameras are becoming increasingly deployed, and therefore this work examines how contextual and vision data obtained from cameras can be integrated with Wi-Fi localization techniques. We present an approach called CALM that works on commodity APs and cameras. Our approach contains several contributions: a camera line fitting technique to restrict the search space of candidate locations, single AP and camera localization via a deprojection scheme inspired from 3D cameras, simple and robust AP weighting that analyzes the context of users via the camera, and a new virtual camera methodology to scale analysis. We motivate our scheme by analyzing real camera and AP topologies from a major vendor. Our evaluation over 9 rooms and 102,300 wireless readings shows CALM can obtain decimeter-level accuracy, improving performance over previous Wi-Fi techniques like FTM by 2.7× and SpotFi by 2.3×.**

***Index Terms*—Wi-Fi, Localization**

I. Introduction

Wireless localization is an important problem, as providing indoor localization is extremely useful across many industry verticals, from indoor navigation [1], to retail consumer analytics [2], to more informed architectural engineering [3]. Wireless localization has gained significant research attention in part because re-using pervasively deployed wireless APs requires no new infrastructure to be installed [4]. Today, however, currently deployed infrastructures commonly feature equipment beyond wireless APs. For example, network-connected cameras have gained increasing prevalence. As many wireless providers include camera offerings in their portfolio [5], [6], [7], it's natural to ask: how can today's infrastructure, consisting of co-deployed APs and cameras, improve localization accuracy?

Retail stores, airports, city buildings, warehouses, and many other buildings deploy cameras for a variety of reasons, ranging from surveillance and security, to analytics, to productivity analysis [8]. Cameras are integrated into the IT infrastructure via wireless or wired links. Today, many cameras feature onboard compute or can connect to an edge computing deployment, enabling advanced analytics to be performed on the live camera feed. Cameras are useful to localization for a variety of reasons. First, the camera "sees" a human or object within its view, giving valuable location information, such as where the user resides in the frame. In addition, cameras also obtain *contextual information* about subjects in the field-of-view, such as their orientation or how they've moved over time. But camera feeds on their own are useless without tools to analyze images. Luckily, advances in deep learning have achieved human-level accuracy in a variety of tasks, such as object detection [9], object tracking [10], pose estimation [11], orientation analysis [12], and more. Armed with such tools, locating subjects and obtaining their context reveals valuable information that can be combined with Wi-Fi readings to enhance localization.

In this paper, we present a Camera-based, AP-integrated Localization Mechanism (CALM) that combines AP and camera readings to improve localization. A key challenge is to combine data intelligently from the RF and visual domains. A noisy reading from a wireless AP can reduce accuracy. CALM combines limited, but highly accurate, location information from monocular cameras with additional Wi-Fi localization estimates to improve accuracy. CALM also utilizes contextual information from cameras to further improve results.

A key design choice of CALM is *simplicity*. We build CALM from known and well-understood techniques in order to lower the bar for deployment. Systems that are easy to understand are easier to manage, debug, and reason about in practice. CALM is flexible: it can work with one or more cameras or one or more APs, cameras need not be colocated with APs, and the techniques presented in this paper can be generalized to improve different types of wireless localization (from angle of arrival, received signal strength, or time-of-flight techniques). All techniques in this paper work with commodity, off-the-shelf cameras and APs, increasing the practicality of the solution. This paper shows the combination of vision and Wi-Fi based localization improves Wi-Fi localization performance. CALM can be thought of as an addition to localization frameworks: when cameras are available CALM will improve performance, but when vision information is unavailable localization can simply fall back to wireless-based techniques. In summary, this paper makes the following contributions:

- A characterization of deployments containing both cameras and APs via a partnership with a major vendor who offers both in their product line.
- A novel mechanism to perform highly-accurate single AP localization (when the AP is colocated with a camera) in

a 3D coordinate space, with no constraints on the number of AP antennas or antenna configuration. Our system takes inspiration from deprojection techniques utilized in 3D cameras.
- A new camera-based trilateration technique to effectively combine limited, but known, location information from a camera with multiple AP readings. The camera contains highly-accurate information about the user's location, which restricts the possible coordinates a user may reside. This approach allows a camera to be located at an arbitrary position.
- A simple and robust technique to weight different AP readings by analyzing the context of the user in the camera frame. A user's orientation is used to intelligently combine readings from wireless APs.
- A codebase, dataset, and evaluation that will be open-sourced. Both 2D and 3D measurements are obtained over 9 rooms, halls, labs, and cafes with 102,300 total wireless readings. Average localization accuracy improves over techniques like FTM [13] and SpotFi [14] by 2.7× to 2.3× respectively, with some rooms showing up to 7.8× improvement.

II. Background

This section briefly provides background to our work. First, various wireless localization techniques are discussed. Then camera-based localization is overviewed. Finally, motivation describing Wi-Fi and camera co-deployments is provided. A more thorough analysis of related work is addressed in Section V.

Wireless localization

Wireless localization can employ a variety of techniques. For example, early approaches use received signal strength (RSS) [4] to determine positioning but are typically prone to high error because RSS is impacted by multipath and attenuation. More recent approaches use an array of antennas to obtain the angle-of-arrival (AoA) of a client's signal [14]. When the client's signal is obtained from multiple APs, a location estimate can be inferred via triangulation. AoA techniques place requirements on antenna configurations and are also impacted by multipath: signals may bounce off walls or other objects instead of directly traversing from the client to the AP. Finally, other approaches utilize time-of-flight, such as 802.11 Fine Time Measurement (FTM), to measure RTTs and infer distance from an AP [13]. With multiple AP readings, trilateration can be used to pinpoint a client's location.

This work mostly focuses on Wi-Fi localization provided by FTM. FTM is standardized in 802.11mc, supported in Android 9 (and above), and is implemented by many device manufacturers [15]. Just as with previous localization techniques, FTM performance can degrade in multipath settings [16], [17], and thus multipath must be addressed when designing an accurate technique. This paper primarily studies CALM with FTM, but such coupling is not fundamental: Section IV presents results with CALM integrated with AoA-based techniques.

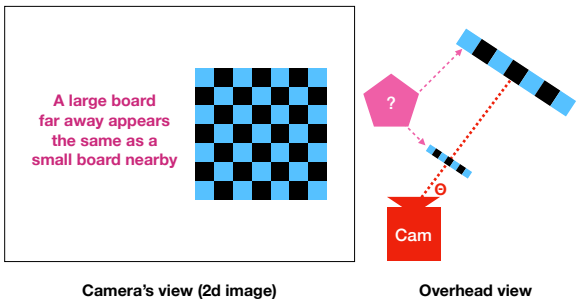


Fig. 1: Single camera lens ambiguity.

Camera localization

Today, most commercially available off-the-shelf cameras contain a single lens (*i.e.,* they are monocular cameras) [18] and hence capture a 2-dimensional image. Consider the checkerboard in Figure 1. One cannot infer if a small checkerboard is placed near the camera or if a large checkerboard is placed farther away. Hence, an object's depth information cannot be inferred and a location estimate is difficult to provide. One can, however, accurately infer the *offset* of the checkerboard in the image (the red dotted line), in both the vertical and horizontal directions.

Obtaining the offset for the checkerboard requires locating the checkerboard in the image. Assuming human subjects instead of checkerboards, an algorithm must be available to locate the human within the 2D image. If the human is located, the centroid of the human's outline can serve as the offsets on the image. Thankfully, object detection and segmentation techniques that utilize deep learning can perform this task quickly and accurately [9]. In Section III, we show how this information helps infer location, but the main idea is the object must appear along the offsets – the challenge is finding the depth. Because object detection techniques are highly accurate, the offset information is close to ground truth, and thus important information regarding an object's location can be captured with high confidence.

Finally, depth information can be gleaned from cameras with multiple lenses. These 3D cameras estimate the depth of an object in an image and then output a location estimate using offset and depth values. However, most cameras deployed today do not support 3D technology, and commodity depth cameras typically have short ranges within 8-10 meters [19]. CALM mostly focuses on monocular camera integration, but we believe our scheme can also improve depth camera localization (*e.g.,* by extending the range of many depth cameras).

Co-location of Cameras and Wi-Fi Our work advocates for co-utilizing cameras and APs in order to improve localization. Several wireless AP enterprise vendors [5], [6], [7] provide *wireless-capable* smart cameras, which allows cameras to seamlessly deploy into wireless networks and could also easily enable smart cameras with integrated Wi-Fi AP functionality in the future. We partner with a major enterprise vendor that includes both Wi-Fi cameras and Wi-Fi APs in their portfolio and sample 1,000 networks containing at least eight cameras

and eight APs. Figure 2 shows a CDF of the ratio of APs to cameras in these networks. A ratio of 1:1 implies equal number of cameras and APs, and a ratio of 2:1 indicates APs are twice as numerous as cameras in a given deployment. For small networks (minimum 16 devices) the median ratio is 1.3, which increases to 1.92 for medium-sized networks (minimum 50 devices) and 2.6 for large networks (minimum 100 devices). The findings have numerous ramifications. First, this suggests the co-location of APs and cameras is common in enterprise deployments. Second, the median ratios are relatively low (especially in smaller deployments), which motivates the need for localization techniques that require only a single AP and single camera. Third, there is a non-trivial amount of networks that do have multiple APs deployed per camera, which is more common in larger, more important customers. We use these findings to motivate the design of CALM, which can work with a single AP and camera, or multiple APs and a camera.

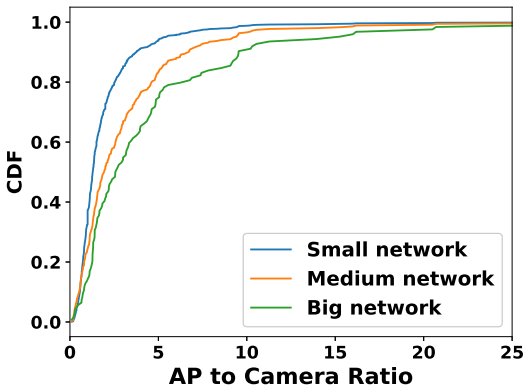


Fig. 2: Co-location of APs and cameras.

III. Design

This section overviews the design of CALM. First, an algorithm is provided when there is a single AP colocated with a single camera. Next, the problem is generalized when multiple APs exist and the camera can be placed in an arbitrary position. Finally, an optimization is provided that uses context derived from the camera's image to obtain a user's orientation and further improve localization.

A. Single AP and camera

When an AP and a camera are colocated, information from each piece of equipment can be combined to unambiguously infer an object's location. The image a camera captures consists of pixels. As seen in Figure 3, a specific pixel is labeled within a *pixel coordinate space* as (u, v), where the origin pixel $(0, 0)$ is in the upper-left corner of the image. An image captures a scene from the physical world, and thus pixels within the image can be mapped to points in a 3D space. These points, measured in meters, are represented as $P = (X, Y, Z)$. Cameras are in part defined by a set of known intrinsic parameters, such as focal length, which is defined in

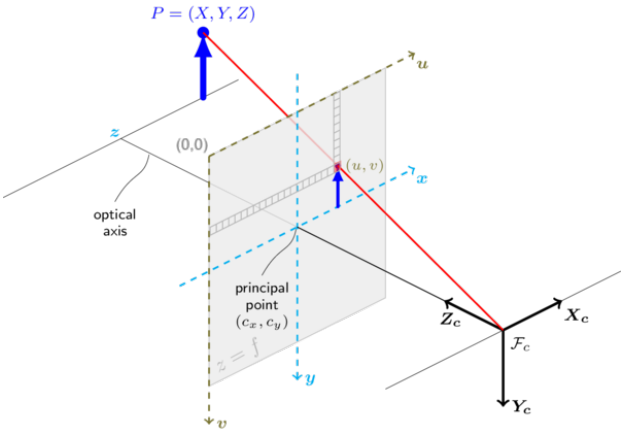


Fig. 3: Deprojection (image derived from [20]).

Fig. 4: Human detection in the camera's view.

pixels for width (f_x) and height (f_y). A known principal point (c_x, c_y) lists the center of projection in pixel coordinates.

In 3D image processing, *deprojection* takes a point within the camera's image (in pixel coordinates) and a depth (in meters) as input and outputs the pixel's location on the 3D physical space. It follows a simple system of equations (where τ is the depth of pixel, denoted by the red line in Figure 3):

$$a = (u - c_x)/f_x \tag{1}$$

$$b = (v - c_y)/f_y \tag{2}$$

$$X = \tau \times a \tag{3}$$

$$Y = \tau \times b \tag{4}$$

The calculation of Z follows in a straight-forward manner once X, Y, and τ are known (*i.e.,* derive Z using the distance formula from the origin). Most cameras deployed today are monocular, which means depth cannot be inferred from the camera alone. When an 802.11 FTM AP is colocated with a monocular camera, however, the depth τ can be derived from the AP's FTM distancing measurement. By combining AP and camera information, a single AP colocated with a single monocular camera can infer the 3D position of an intended wireless object in the camera's view.

To determine the pixel whose depth should be acquired, object detection techniques can be employed. Shown in Figure 4, object detection (or more specifically human detection) finds

the presence of an object (human) in an image and draws a bounding box around the object (human). Bounding boxes are typically rectangular, although more advanced segmentation techniques can also be employed [21]. In CALM, the center of a bounding box (*i.e.,* the approximate center of a human) is used as the pixel in which depth is applied.

B. Multiple APs and camera

Many Wi-Fi localization techniques utilize measurement data from multiple access points since there is usually overlapping coverage in production deployments. CALM can also combine readings from multiple APs to infer location. Before describing how CALM integrates with multiple APs, traditional Wi-Fi FTM-based localization is first described.

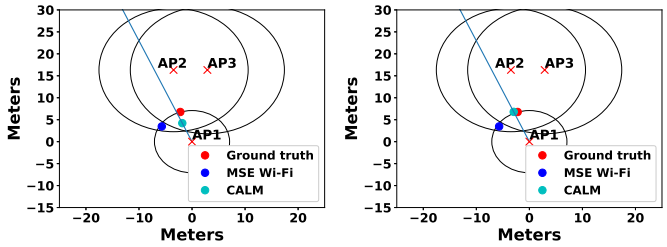
With distance measurements from at least three APs, trilateration techniques can be employed to determine the physical coordinates of an object. Trilateration algorithms search a set of candidate points in physical space, calculating the mean-squared error (MSE) from the measured FTM distance reading to the candidate point's actual distance for each AP. The algorithm outputs the candidate point with lowest MSE. For example, to obtain a user's location via trilateration using Wi-Fi FTM readings, the following equation can be used:

$$P_t = \arg\min\left(\frac{1}{N}\sum_{1}^{N}(\tau_i - \hat{\tau}_i)^2\right) \quad (5)$$

where N is the number of APs, τ_i is the measured distance from AP_i obtained via FTM readings, and $\hat{\tau}_i$ is the candidate point under consideration's distance to AP_i. Optimization algorithms such as Limited-memory BFGS [22] determine the candidate points that are searched, and P_t is the output coordinate that minimizes MSE over all searched points.

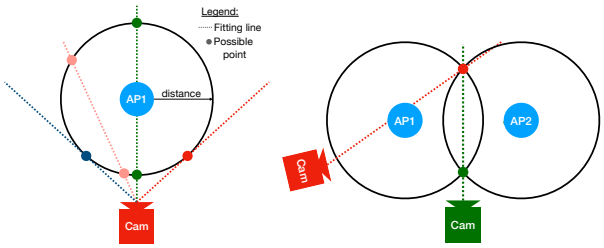
With multiple APs in CALM, the MSE search can be modified because the camera gives important information about the object's location without ambiguity. Say a specific object is located in the pixel coordinate space at (u, v), but the depth of the object is unknown. This is true for all monocular cameras. By considering the object can be *any* depth away from the camera, the physical space in which the object resides must fit on a single straight line. We call this line the *camera fitting line*, which is the red line in Figure 3. Armed with this insight, we modify the MSE search with an extra constraint: only points along the camera fitting line should be considered. In detail, the first candidate point is set to a depth of 0.1m on the camera fitting line and its physical space coordinate is calculated from equations 3 and 4. To search along the camera fitting line, the depth of each candidate point is increased at 0.1m increments. As in equation 5, the candidate point that minimizes the distancing error over all APs is selected as the final output point.

Figure 5a visualizes FTM Wi-Fi and CALM-based localization for the measurement taken in Figure 4. There are three APs in the figure (each denoted by a red ×), a camera colocated with AP_1 at (0,0), and the ground truth location of the user at the red dot (•). The black circles' radii represent the distance to the object as measured by each APs' FTM



(a) CALM-3AP (b) CALM-context

Fig. 5: Localization with CALM.



(a) Single AP (b) Two APs

Fig. 6: Arbitrary camera positioning in CALM.

reading. With FTM Wi-Fi trilateration (denoted MSE Wi-Fi, dark blue dot •), the projected location is 4.79m away from the ground truth, mostly because the APs near the back of the room (AP_2 and AP_3) over-estimate their distance. The camera fitting line (blue line), however, ensures the MSE search is only performed on points within the line, and hence CALM-based measurement gives an error of 2.51m (cyan dot •).

In Section III-A, the camera was required to be colocated with a single AP. With multiple APs, the camera can be more flexibly placed in the topology. With a single AP and a camera at an arbitrary position, searching along the camera fitting line may yield two points with minimized MSE, as shown in Figure 6a. Here, only users placed on the blue and red lines produce unambiguous results, whereas users placed on all other possible camera fitting lines would intersect with the circle twice, leaving two possible user locations. By deploying more APs, positioning the camera away from the APs eliminates most of the ambiguous cases. For example, the red line in Figure 6b shows an unambiguous result. Note some AP configurations may still result in issues, such as when all APs are in a line (here the dotted green line provides an ambiguous result). But such configurations also cause issues with the FTM-based Wi-Fi scheme. The camera may help disambiguate the dotted green line case in Figure 6b by perhaps examining the relative size of the user in the image, but such optimizations are left for future work.

C. Context-aware AP weighting

In addition to providing the camera fitting line, the camera reveals useful contextual data to the localization scheme. One interesting enhancement is to consider the orientation of the user with respect to the deployed APs. Numerous studies have

shown the human body can attenuate signals when the body is between a mobile device and an AP [16], [17]. These non-line-of-sight (NLoS) scenarios can force wireless signals to travel farther distances via a reflected path, which in turn causes higher inaccuracies with FTM ranging. In contrast, when users are facing an AP with clear line-of-sight (LoS), a direct path is more likely, which results in more accurate distance ranging. Indeed, this can be seen in Figure 5a, where the user is facing AP_1 but not facing the other APs (AP_2 and AP_3). The distancing measurement from AP_1 is much more accurate than the distancing measurement from the non-facing APs, and worse yet the non-facing AP error causes Wi-Fi trilateration techniques to perform poorly.

As a result, CALM can use the camera to analyze the orientation of the user and infer which APs the user is facing. The computer vision community has already designed many highly accurate user orientation techniques [12], [23] that can be co-opted for this purpose. At a high level, CALM applies weights to an AP's distancing estimate based on whether the user is facing the AP. The scheme first estimates a user's location using all APs with equal weights (in other words, simply performing the steps outlined in Section III-B). This gives an initial location estimate. Next, the camera image is fed to the user orientation neural network, which outputs the orientation. From there, two lists of APs are created: AP_{LS} in which the user is likely facing those APs and AP_{NS} in which the user is not facing those APs. Then, the MSE search in equation 5 can be modified:

$$P_t = \arg\min\left(\frac{1}{N}\left(\sum_{1}^{AP_{NS}} w_{NS}(\tau_i - \hat{\tau}_i)^2 + \sum_{1}^{AP_{LS}} w_{LS}(\tau_i - \hat{\tau}_i)^2\right)\right) \tag{6}$$

where w_{NS} and w_{LS} are the weights applied to NLoS and LoS readings, respectively. Larger weights should be applied to LoS readings since they are likely to be more accurate than NLoS readings. Just as before, only potential locations on the camera fitting line are considered.

While seemingly simple, the scheme is effective. Figure 5b shows the output of this scheme, denoted CALM-context (cyan dot •), from the ongoing Figure 4 example. CALM-context obtains an error of only 0.33m, outperforming the Wi-Fi baseline by over 14×. In the future, the technique could be optimized by adjusting the weights with more sophistication, such as additionally including distance from the AP as part of the weighting, analyzing the environment in a more complex fashion (*i.e.,* explicitly detecting reflectors or columns in the environment that may block an AP a user is facing [24]), or using advanced techniques to infer wireless signal characteristics from an image (as studied in [25]).

IV. Evaluation

This section first details the methodology of our approach, followed by a description of the evaluation results.

A. Methodology

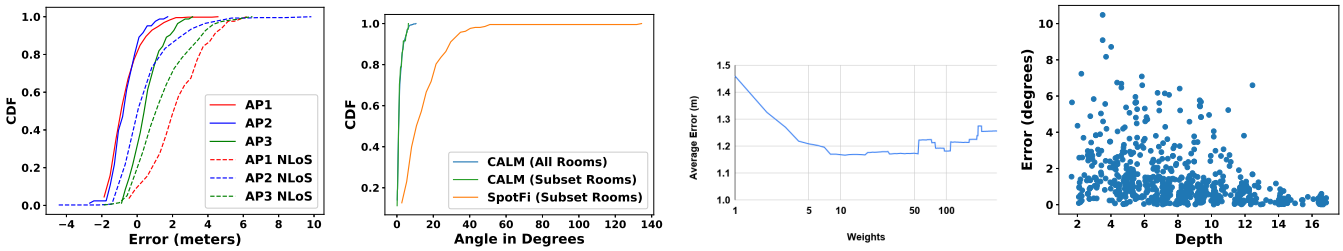
Our experiments utilize commercially off-the-shelf equipment. An Intel RealSense D435i depth camera is used as the camera in our setup [19]. The D435i camera can obtain 3D coordinates with a range up to about 10 meters, depending on lighting, scene, and calibration. When using the camera for CALM's evaluation, we *do not* use the 3D information provided by the camera, instead using the output of the monocular RGB camera. There exists up to three Google WiFi APs, which support FTM measurement. One of the APs is colocated with the camera (denoted AP_1) and the other two APs (denoted AP_2 and AP_3) are typically positioned 10-20 meters from the first AP. We utilize Google Pixel 3 phones as the client, and run the WifiRttScan Android app [26] to collect the FTM measurements. The channel width is configured to 80 MHz. For each AP, a burst of 50 FTM measurements is collected and the minimum of the measurements is used for trilateration (this appears to give the best performance for legacy Wi-Fi trilateration). In addition to FTM, we also test our approach against SpotFi [14]. Three Qualcomm IPQ8076A chipset-based APs using 4 antennas obtain CSI data, with each AP positioned similar to its FTM-based counterpart. SpotFi methodology is taken from [14]: channel width is 80MHz, AoA readings are taken over 10 packets, and a clustering algorithm is used to find the AoA of the highest-likelihood direct path. We compare the following approaches:

- **Ground truth** The actual coordinates obtained with a high-precision laser ranging device.
- **Camera** We obtain the 3D location estimates from the Intel D435i SDK.
- **Wi-Fi** A trilateration approach, in which FTM measurements are collected from all three APs and the MSE dictates final location.
- **SpotFi** A widely-cited localization approach that minimizes AoA and RSSI-based distance errors.
- **CALM-1AP** Our approach using FTM readings from a single AP (AP_1), which is co-located with the camera.
- **CALM-3AP** Our approach searching the MSE using the camera's fitting line over all three APs (Section III-B).
- **CALM-context** User orientation gleaned from the camera is used to weight each AP (Section III-C).

Room	Size (m)	# pts	AP_2 loc	AP_3 loc	3D	Or
Conference	8.7 x 19.6	41	-3.4, 16.3	2.8, 16.3	N	1
Class1	11.8 x 12.6	72	-5, 11.2, 1.8	4.5, 11.2, 1.8	Y	1
Class2	9.1 x 8.8	26	-3.4, 6.5, 0.5	3, 6.5, 0.5	Y	1
Class3	11.4 x 13.2	50	-6.6, 11.6, 1.5	2.3, 11.6, 1.5	Y	1
Class4	8.8 x 12.2	49	-4.2, 9.6	4.2, 9.6	N	1
Study hall	11.2 x 23.5	49	-4.5, 17.2	4.1, 17.2	N	1
ML Lab	4.6 x 12.1	23	-1.3, 11.5	3.2, 11.5	N	1
Cafe	8.4 x 13.6	54	-6.3, 13.6	4.4, 13.6	N	1
Systems Lab	15.3 x 20.8	32	-5.3, 14.6	5.3, 14.6	N	1
Conference	8.7 x 19.6	37	-3.7, 16.3	2.5, 16.3	N	2
Class2	9.1 x 8.8	15	-3.5, 6.4, 0.5	3.3, 6.4, 0.5	Y	2
Class3	11.4 x 13.2	59	-6.5, 11.2, 1.5	2.4, 11.2, 1.5	Y	2
Systems Lab	15.3 x 20.8	32	-5.3, 14.6	5.3, 14.6	N	2

TABLE I: Experimental information. AP_1 is located at the origin. Or = Orientations.

Table I details the rooms used in our experiments. For each room, the following are listed: the room size, the number of points considered in each room, AP_2 and AP_3 locations, whether the points we localize are on a 2D or 3D plane,



(a) Errors in AP distancing (b) CDF of errors in AoA (c) Weight sensitivity (w_{LS}) (d) AoA depth error histogram

Fig. 7: Microbenchmarks.

and how many orientations the user positions during the test. We break the dataset into two categories: single orientation results and multiple orientation results. For single orientation results, the user is always facing the camera and the colocated AP_1. For multiple orientation results, the user either directly faces the camera or faces the opposite direction. In total, we analyze over 102,300 wireless readings from a variety of rooms such as classrooms, labs, cafes, study areas, and conference rooms. Comparison against SpotFi required two sets of APs to be deployed, and thus experiments were performed over a subset of the rooms. From Table I, `Class4` (shown in Figure 12a) was chosen as a representative room and `Conference` (Figure 4) was chosen for a challenging room with significant multi-path.

Techniques like CALM and SpotFi must know the orientation of the cameras and APs to accurately infer user locations. With CALM, the center of the camera could be aligned with a known point on an opposite wall to obtain ground truth reference points. SpotFi and CALM could also obtain ground truth orientation by collecting a few points at known locations and then adjusting AP and camera orientation to minimize error. The latter technique is performed in this paper.

B. Microbenchmarks

This section provides a number of microbenchmarks to help understand the performance of our scheme.

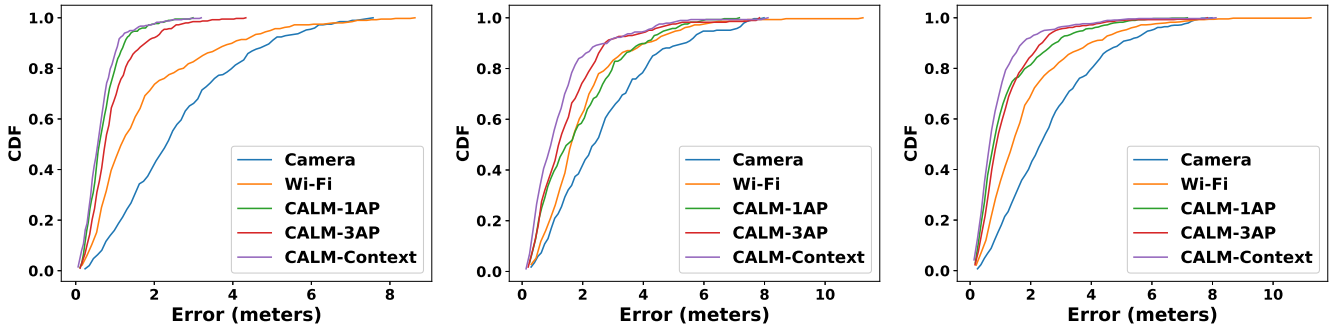
Equipment calibration and facing validation Similar to other studies [17], we find FTM readings can sometimes under-report distances to a device. Figure 7a shows a CDF of the distancing error from each AP to the wireless client when the human is facing the AP. This data is collected from our multiple orientation dataset. We find (i) APs can under-report distances in these cases and (ii) different APs may have different median errors. Because CALM relies on accurate FTM measurements, especially with a user facing an AP, we use the median error as a fixed offset for every AP's reading in our experiments. Note this offset only needs to be measured once. For the Wi-Fi baseline, we use whichever is best (using the offset versus not using the offset).

The graph also shows the CDF of ranging errors when the user is not facing an AP. NLoS ranging measurements overestimate distances compared to LoS ranging measurements. The graph verifies this trend and justifies our reasoning to weight facing APs more heavily in CALM-context. Note as opposed to the facing FTM readings, we do not apply NLoS offsets to the FTM readings since such offsets are likely a product of the environment. We leave as future work to study how to incorporate NLoS offsets in more detail.

Orientation accuracy We utilize WHENet [12] to obtain user orientation data. Camera images are fed into WHENet, which outputs an orientation and allows us to determine which APs the user is facing. We *do not* provide any of our images for training the network nor do we run any type of transfer learning– we simply run the network off-the-shelf. The average accuracy over the whole dataset is 0.97, with the lowest accuracy being 0.88 (Conference room). Over our whole dataset, using WHENet introduces roughly 5.5% localization error when compared to an oracle-based scheme that knows ground-truth orientations.

Angle of arrival (AoA) error CALM's camera fitting line essentially provides an angle of arrival (*i.e.,* the angle in which the user is offset from the center of the camera's projection). Figure 7b shows a CDF of the error, defined as the absolute value of the difference between the actual angle and the measured angle. The median error for CALM over all rooms is 0.96 degrees. For the subset of rooms with SpotFi measurements, CALM's median error is 0.90 degrees while SpotFi in same rooms has a median AoA error of 10.73. Hence, camera-based AoA derivation significantly outperforms SpotFi. We were originally quite surprised that some of the camera's angle errors are relatively high (6-10 degrees) for CALM, but a careful analysis of the data showed these points typically occur in two cases. First, when users are close to the camera, small errors in object detection manifest themselves as large angle errors. Figure 7d shows the error is much higher when the object is at a depth closer to the camera. Second, when a user appears near the edge of the camera's field-of-view, distortion can introduce error [27]. It may be possible to apply anti-distortion techniques to mitigate these errors, but we find overall performance to still be reasonable and hence leave such optimizations to future work.

Weight sensitivity CALM-context weights readings from APs the user is facing more heavily than non-facing AP readings. Figure 7c shows the average error, over the multiple orientation dataset, of CALM-context with varying facing weights (non-



(a) Single orientation rooms (b) Multiple orientation rooms (c) Combined orientation rooms

Fig. 8: CDF of errors over all rooms for different orientations.

Room	Camera	Wi-Fi	CALM-1AP	CALM-3AP	CALM-context
Conference	3.637	3.538	1.011	1.752	0.937
Class1	2.313	0.974	0.664	0.635	0.640
Class2	1.060	1.101	0.540	0.728	0.544
Class3	2.486	1.436	0.651	0.956	0.734
Class4	1.728	1.146	0.744	0.714	0.719
Study hall	3.917	1.352	0.490	0.599	0.457
ML Lab	1.634	1.807	0.613	0.791	0.575
Cafe	2.793	2.900	0.371	0.875	0.436
Systems Lab	2.563	1.251	0.675	0.625	0.629
Average	2.459	1.723	0.640	0.853	0.630

TABLE II: Average error (in meters) for single orientation rooms. Green highlights the best performer per room.

Room	Camera	Wi-Fi	CALM-1AP	CALM-3AP	CALM-context
Conference	4.175	3.445	3.006	2.282	2.070
Systems Lab	2.483	1.654	1.597	1.265	0.975
Class2	1.065	0.960	1.174	0.925	0.607
Class3	2.194	1.558	1.388	1.273	1.065
Average	2.479	1.904	1.791	1.436	1.180

TABLE III: Average error (in meters) for multiple orientation rooms. Green highlights the best performer per room.

facing weight w_{NS} is set to 1). The scheme is robust to a variety of weights, with only a small difference in performance based on weights in the range of 10 to 100. In our experiments, we set the weight to 11.5. More advanced techniques in the future may set the weight on a per-room basis (we did see benefits for such an approach), but we aim for simplicity in deployment and find a single weight works well.

Pipeline runtime It takes 0.73s to run the CALM pipeline on a laptop with an 8^{th} Gen Intel Core i5-8265U CPU and 8GB DDR4 memory (0.65s for YOLOv3, 0.07s for WHENet, 0.014s for additional CALM logic). YOLOv3 is run on the CPU, but latencies can be as low as 30ms if run on a GPU [9].

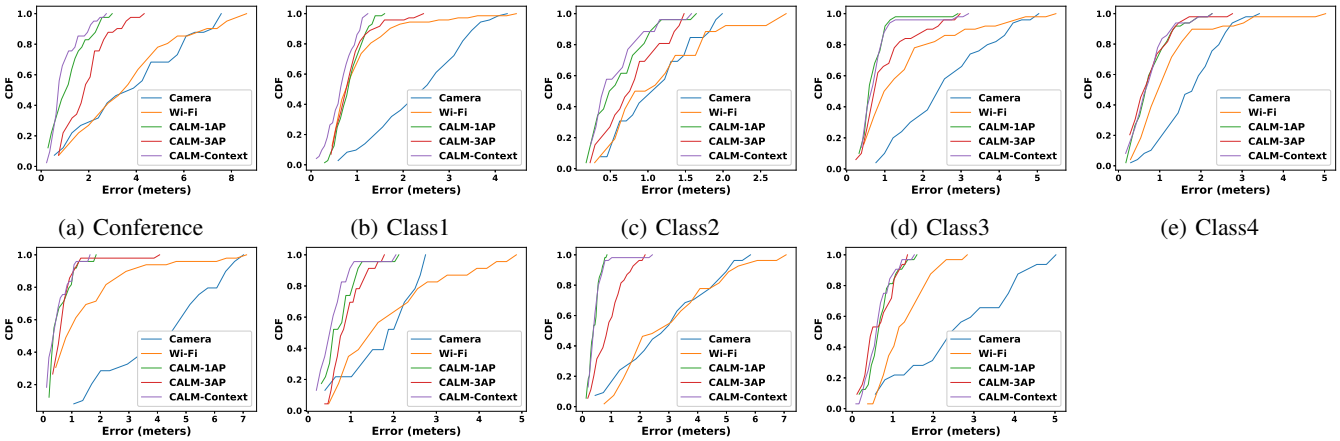
C. Localization Results

This subsection analyzes localization performance. First, a deep dive of CALM's performance is presented, with results from single and multiple orientation datasets discussed. Afterward, CALM is compared to SpotFi on a subset of the dataset.

1) Single Orientation Results: Single orientation results span 9 rooms (Table I), the results are analyzed over a series of graphs. Figure 8a shows a CDF of location error over all single orientation rooms for Camera, Wi-Fi, CALM-1AP, CALM-3AP, and CALM-context. Table II shows the average error for each scheme, broken down on a per-room basis. Figure 9 shows the CDF of errors for each room within the dataset.

A number of trends can be noted from the data. First, CALM-1AP and CALM-context perform admirably, with average error rates of 0.64m and 0.63m, respectively. This represents a 2.7× increase over Wi-Fi error (1.72m) and a 3.9× increase over Camera error (2.46m). CALM-1AP outperforms CALM-3AP (0.85m average error) because the user is facing AP_1 (the single AP), which typically gives accurate distancing information. When two other APs are involved, in which the user is not facing, the average error increases because the two APs add noise to the localization process. CALM-context also uses three APs, but instead intelligently weights AP_1's measurements higher than AP_2 and AP_3's measurements. As a result, CALM-context is able to closely track or beat CALM-1AP's performance in many cases. On a per-room basis, CALM-1AP and CALM-context show gains over Wi-Fi ranging from 1.4-7.8× and 1.5-6.6×, respectively. The per-room CDFs show the CALM techniques perform well, with the camera performing relatively better on the smaller rooms.

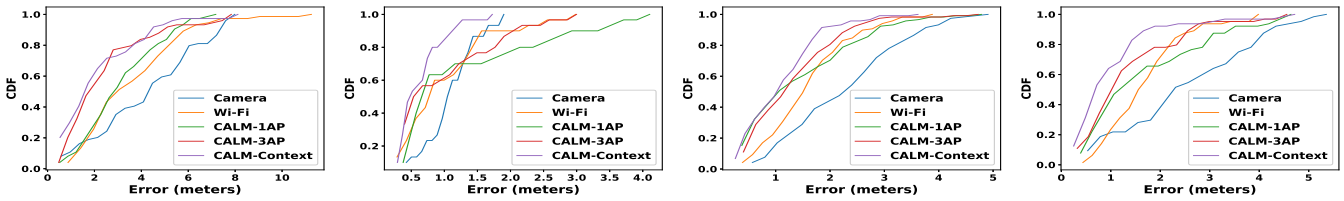
2) Multiple Orientation Results: Next, results are analyzed when the user faces AP_1 and then also turns 180° at each location. Figure 8b shows the CDF of location errors of each scheme over all four multiple orientation rooms. Table III shows the average error of each approach, broken down per room. Figure 10 shows the CDFs for each room. While CALM-3AP and CALM-context again both outperform Wi-Fi (average improvements of 1.3× and 1.6×, respectively), different from the single orientation results, CALM-1AP gain's are reduced (1.06×). In the single orientation dataset, the user was always facing AP_1. However, with multiple orientations, CALM-1AP can perform poorly when the user is not facing the AP colocated with the camera (*i.e.,* AP_1). One should note, however, that CALM-1AP, which requires one AP, is able to achieve similar levels of accuracy as the baseline Wi-Fi, which requires three APs. On a per-room basis, CALM-context's gains over Wi-Fi range from 1.5-1.7×.



(a) Conference (b) Class1 (c) Class2 (d) Class3 (e) Class4

(f) Study Hall (g) ML Lab (h) Cafe (i) Systems Lab

Fig. 9: CDF of errors for each single orientation room.



(a) Conference (b) Class2 (c) Class3 (d) Systems Lab

Fig. 10: CDF of errors for each multiple orientation room.

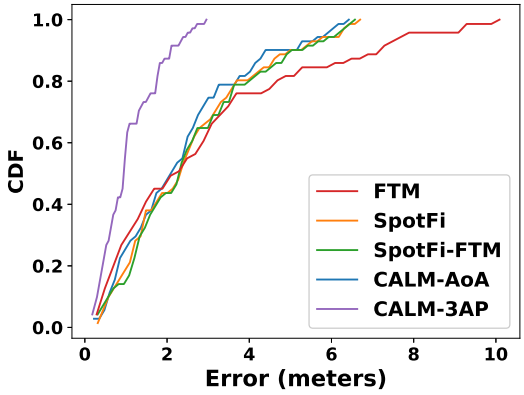


Fig. 11: CDF of location Errors with SpotFi.

(a) SpotFi Experiment (b) Pairing Experiment

Fig. 12: Rooms for experiments.

3) Combined Orientation Results: Next, we analyze over the whole dataset. Figure 8c shows a CDF of each scheme over all rooms. The average improvements of CALM-1AP, CALM-3AP, and CALM-context over Wi-Fi are 1.83×, 1.76×, and 2.25×, respectively.

4) SpotFi Comparison: Finally, we compare to SpotFi on a subset of the single orientation rooms: `Conference` as pictured in Figure 4 and `Class4` as shown in Figure 12a. Figure 11 shows the location error over both rooms for Wi-Fi FTM, SpotFi, CALM-3AP, and two new techniques. *SpotFi-FTM* uses FTM instead of RSSI to estimate distance from the APs and *CALM-AoA* uses AoA from the AP2 and AP3 instead of FTM. CALM-AoA still uses the camera fitting line technique outlined in Section III-B and hence shows how CALM can be integrated into AoA-based techniques. Note the CALM techniques do not use the user orientation optimization. Table IV shows the average errors. The results indicate a number of trends. First, FTM typically outperforms AoA (*i.e.,* SpotFi) on our dataset. SpotFi's errors mostly come from incorrect AoA assessment, and hence SpotFi-FTM and SpotFi (which uses RSSI instead of FTM) closely track one another. CALM-AoA improves either SpotFi technique by forcing location along the highly-accurate camera-fitting line: average errors reduce from 2.5m to 2.3m. CALM-3AP improves CALM-AoA by 2.1× on average because AoA has higher errors in our testing environment. Last, CALM-3AP improves SpotFi by 2.3× on average.

Room	SpotFi	FTM	SpotFi-FTM	CALM-AoA	CALM-3AP
Conference	2.605	4.589	2.7332	2.446	1.217
Class4	2.411	1.148	2.372	2.198	0.916
Average	2.508	2.868	2.552	2.322	1.066

TABLE IV: Average error (in meters) for SpotFi comparison. Green highlights the best performer per room.

Camera Position	Wi-Fi	CALM-1AP	CALM-3AP	CALM-context
VC1	1.904	1.746	1.422	1.068
VC2	1.904	0.896	1.026	0.889
VC3	1.904	1.157	1.044	0.990
VC23	1.904	-	1.238	1.050

TABLE V: Average error (in meters) over multiple orientation rooms with virtual cameras. Green highlights the best performer per room.

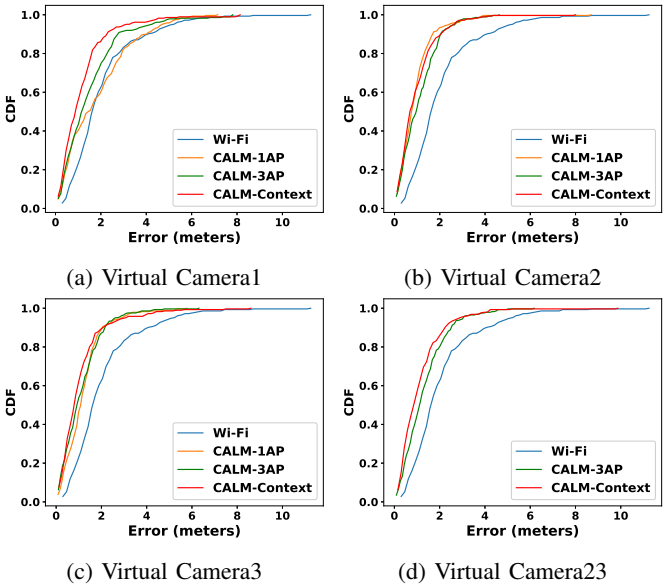
D. Virtual Camera Results

This section introduces a novel evaluation methodology termed the *virtual camera*, where the camera is virtualized and placed at arbitrary locations within each room. Because the human can often be detected in an accurate manner via the camera, the angle-of-arrival error is typically small in CALM (Figure 7b). With virtual camera, the error is assumed to be zero, and hence the camera fitting line can be drawn from the virtual camera through the human's known location. User orientation is also assumed correct via an oracle. Each AP's FTM measurements can be reused from the previous experiments to provide a location estimate. Therefore, virtual camera results give a lower-bound on CALM localization error, but also give the flexibility to perform "what-if" analysis of moving the camera to different positions in the room.

Figure 13 shows CDFs over all multiple orientation rooms for different virtual cameras, and Table V shows the average location errors. VCi indicates the camera is located at AP_i, and VC23 is when the virtual camera is placed half-way between AP_2 and AP_3 (note CALM-1AP cannot be evaluated in the VC23 case because it requires colocation to an AP). The results show CALM works effectively regardless of camera position. Interestingly, the last row (Average) in Table III can be directly compared to the VC1 result in Table V because the camera is located at AP_1 in each case. The schemes that do not utilize orientation, CALM-1AP and CALM-3AP, perform negligibly worse without the virtual camera, showing CALM's mechanism to generate the camera fitting line adds little error to overall accuracy. CALM-context sees about 10% degradation, which mostly comes from orientation estimates being off. Overall, the virtual camera results show CALM-context can increase accuracy over Wi-Fi by 1.78-2.14×.

E. Multiple Users

Our previous experiments located a single user. With multiple users the wireless device in the RF domain must be *paired* with its associated user in the visual domain. Our study mostly focuses on how accurate camera and Wi-Fi localization can be, and the pairing issue can be addressed by previous work [28], [29], [30], [27], [31]. Such schemes pair movement patterns of users from the Wi-Fi and vision domains and even work



(a) Virtual Camera1 (b) Virtual Camera2

(c) Virtual Camera3 (d) Virtual Camera23

Fig. 13: CDF of errors for Virtual Cameras over all multiple orientation rooms.

in cases when the number of users does not equal the number of devices. Prior pairing works report high accuracies, ranging from 87-95% [28], 84-94% [30], 95% [29], and 90% [31].

Regardless, we also test the accuracy of pairing within our testbed, as shown in Figure 12b with four users in `Class3`. Here, users randomly position themselves in the room over 25 different trials. We locate Wi-Fi devices using the FTM technique and estimate the users' physical locations from the image via a two-step process. First, YoLo detects human bounding boxes. Second, the pixel within the center of the bounding box is projected along the camera fitting line by estimating the depth of the user from the size of a bounding box. In a separate room, we previously trained a linear model relating the size of a bounding box to the user's depth. This linear model is used in testing to translate bounding box sizes into depth and obtain physical coordinates via the equations in Section III-A. Running the Hungarian algorithm [32] on these two sets of points yields 78% pairing accuracy. We believe accuracy can be further improved by analyzing user trajectories and by locating cameras near the ceilings– in our scheme cameras near waist-height often leave users overlapping as seen in Figure 12b.

V. Related Work

Related work covers Wi-Fi localization, Wi-Fi localization with cameras, infrastructure-based localization, image processing, and vision-based localization.

Wi-Fi localization: Radar [4] is a seminal paper in Wi-Fi localization, using received signal strength (RSS) to localize users. Since then, research in Wi-Fi localization has been vast. Techniques have ranged from using RSS, angle-of-arrival, multipath, backscatter, various frequencies and time-of-flight [33]. CALM is complementary to the large body of previous

research. Much of the research aims to overcome localization issues due to multipath, reflections, or varying and dynamic wireless conditions. As wireless techniques improve, they can be integrated into CALM's camera-based schemes.

Wi-Fi localization with cameras Some localization techniques perform localization when Wi-Fi is augmented with additional equipment. For example, Irshad *et al.* [27] augment a 3D camera's performance by using Wi-Fi localization when outside the camera's accurate field-of-view. iVR [34] uses multiple cameras, an accelerometer on a user device, and wireless RSS-based localization. CALM does not require extra user participation or equipment, nor does it require multiple cameras. EV-Loc [31] integrates Wi-Fi RSS readings and visual signals for localization by first finding locations in each domain. In the vision domain, an intensive, offline data collection phase captures images of a user at every location. In the wireless domain, path loss models obtain distance estimates from RSS values. After obtaining both location estimations, points from both domains are matched, and the final location is found by weighting points in each domain. EV-Human [35] is an extension of EV-Loc that compensates for human body attenuation of wireless signals by determining the positioning of the user and device via cameras. There exist several differences between EV-Human and CALM. First, EV-Human is RSS-based and utilizes an empirically-based RSS compensation technique, both of which require significant fingerprinting and overhead. CALM has no such overheads and applies a simple and robust weighting scheme to FTM-based APs in order to obtain accurate localization. EV-Human also requires multiple cameras, and multiple nearby APs to be colocated on the same channel, to in part compensate RSS due to orientation. CALM works with a single camera and single AP, and neighboring APs need not be tuned to the same channel. In summary, our work explores novel ideas when compared to previous work: the camera fitting line approach, single AP and camera localization via deprojection, simple and robust context-based AP weighting, and the virtual camera methodology. Last, RGB-W [36] maps wireless signals to images and then utilizes a dictionary-based technique and a cascade of convex solvers to improve localization. RGB-W relies on a series of complex models to generate noise detection estimates, wireless radius ranges, and error bounds. CALM does not require offline data collection, multiple cameras or APs, nor sophisticated wireless models.

Finally, another class of work locates wireless devices to be displayed in augmented reality applications [37]. CALM is likely to help in these scenarios because an augmented reality device already contains a camera.

Image processing There is vast literature on object/human detection and tracking [38], [9], [39]. These schemes achieve high accuracy and are continuously being developed to run efficiently. CALM can judiciously utilize advances in this area to improve performance.

Vision-based localization Stereovision is a common technique utilizing two cameras to measure distance. Some recent works utilize a deep convolutional neural network for computing the depth of an RGB image [40], [41]. These works typically require significant amounts of training data and usually output relative, rather than absolute, depth.

VI. Discussion and Conclusion

This section discusses limitations and opportunities of our work before concluding.

Issues with vision-based optimizations Our work assumes the user is located within the camera's field-of-view. Users behind the camera, too far away, or obstructed by another object cannot explicitly be localized with CALM. A user temporarily obstructed by another object could be tracked over time, however, and these movement patterns could help infer past or current locations. In addition, it may be possible to use the *absence* of a user in the field-of-view for benefit. Such information could be useful for geo-fencing by determining if a user is inside or outside a room. We leave such optimizations as future work.

Vision-based techniques can suffer from issues such as poor lighting. We assume cameras are deployed in high-value areas with sufficient lighting. In summary, CALM is always a net win: when cameras are available, accuracy is improved and when cameras (or viewpoints) are unavailable, legacy Wi-Fi techniques can be employed to maintain previous performance.

Non-human Wi-Fi users While we assume humans are located via their wireless devices, other non-human subjects could also be localized. For example, a robot vacuum could be localized as long as a robot vacuum object detection scheme can be derived. Given image processing techniques can obtain human-level object detection accuracy on ImageNet's [42] thousands of classes, such alterations seem plausible.

Conclusion This paper explores how camera and Wi-Fi infrastructure can be used together to improve localization. Today, cameras with network connectivity are being increasingly deployed, and the cameras are typically outfitted with intelligence or connected to an edge computing framework. As such, we explore how a single AP colocated with a single camera can locate users in 3D physical space, and then extend our work to support multiple APs and cameras at arbitrary positions. Our system, called CALM, contains several novel contributions: a camera line fitting approach to restrict the search space of candidate locations, single AP and camera localization via a deprojection technique inspired from 3D cameras, simple and robust AP weighting that analyzes the context of users via the camera, and a new virtual camera methodology. We partner with a major wireless and camera vendor to analyze today's deployments and further motivate our scheme. Our evaluation over 9 rooms and 102,300 wireless readings shows CALM can obtain decimeter-level accuracy, improving performance over Wi-Fi techniques like FTM by 2.7× and SpotFi by 2.3×.

Acknowledgements This work is partially funded by NSF-1908910.

REFERENCES

[1] J. Biswas and M. Veloso, "Wifi localization and navigation for autonomous indoor mobile robots," in *2010 IEEE International Conference on Robotics and Automation*, 2010, pp. 4379–4384.

[2] R. Nandakumar, S. Rallapalli, K. Chintalapudi, V. Padmanabhan, L. Qiu, A. Ganesan, S. Guha, D. Aggarwal, and A. Goenka, "Physical analytics: A new frontier for (indoor) location research," Tech. Rep. MSR-TR-2013-107, October 2013, microsoft Research Technical Report. [Online]. Available: https://www.microsoft.com/en-us/research/publication/physical-analytics-a-new-frontier-for-indoor-location-research/

[3] R. Melfi, B. Rosenblum, B. Nordman, and K. Christensen, "Measuring building occupancy using existing network infrastructure," in *2011 International Green Computing Conference and Workshops*, 2011, pp. 1–8.

[4] V. Bahl and V. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," in *Proceedings of IEEE INFOCOM 2000*, March 2000.

[5] "Cisco meraki cloud managed cameras," https://meraki.cisco.com/products/smart-cameras/.

[6] "Aruba network's video surveillance solution," https://www.arubanetworks.com/pdf/solutions/AB_VIDSUR.pdf.

[7] "Google nest," https://store.google.com/us/category/google_nest.

[8] "Global smart cameras market forecast," https://www.marketresearchfuture.com/reports/smart-cameras-market-1326.

[9] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

[10] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *arXiv preprint arXiv:2004.01888*, 2020.

[11] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt, "Xnect: Real-time multi-person 3d motion capture with a single rgb camera," *ACM Trans. Graph.*, vol. 39, no. 4, Jul. 2020. [Online]. Available: https://doi.org/10.1145/3386569.3392410

[12] Y. Zhou and J. Gregson, "Whenet: Real-time fine-grained estimation for wide range head pose," *arXiv preprint arXiv:2005.10353v2*, 2020.

[13] R. Want, W. Wang, and S. Chesnutt, "Accurate indoor location for the iot," *Computer*, vol. 51, no. 08, pp. 66–70, aug 2018.

[14] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using wifi," in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, ser. SIGCOMM '15. New York, NY, USA: ACM, 2015, pp. 269–282. [Online]. Available: http://doi.acm.org/10.1145/2785956.2787487

[15] "Android devices that support wifi-rtt," https://developer.android.com/guide/topics/connectivity/wifi-rtt#supported-devices/.

[16] K. Jiokeng, G. Jakllari, A. Tchana, and A. L. Beylot, "When ftm discovered music: Accurate wifi-based ranging in the presence of multipath," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020, pp. 1857–1866.

[17] I. et al., "Verification: Accuracy evaluation of wifi fine time measurements on an open platform," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 417–427. [Online]. Available: https://doi.org/10.1145/3241539.3241555

[18] "3D Camera Market - Forecast(2021 - 2026)," https://www.industryarc.com/Report/15306/3d-camera-market.html.

[19] "Intel d435i," https://www.intelrealsense.com/depth-camera-d435i/.

[20] "Opencv: Camera calibration and 3d reconstruction," https://docs.opencv.org/2.4/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html, journal=OpenCV online documentation.

[21] L. Yang, Q. Song, Z. Wang, and M. Jiang, "Parsing r-cnn for instance-level human analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[22] "Limited-memory bfgs," https://en.wikipedia.org/wiki/Limited-memory_BFGS.

[23] J. Choi, B. Lee, and B. Zhang, "Human body orientation estimation using convolutional neural network," *CoRR*, vol. abs/1609.01984, 2016. [Online]. Available: http://arxiv.org/abs/1609.01984

[24] R. Ayyalasomayajula, A. Arun, C. Wu, S. Sharma, A. R. Sethi, D. Vasisht, and D. Bharadia, *Deep Learning Based Wireless Localization for Indoor Navigation*. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3372224.3380894

[25] S. Ayvaşundefinedk, H. M. Gürsu, and W. Kellerer, "Veni vidi dixi: Reliable wireless communication with depth images," in *Proceedings of the 15th International Conference on Emerging Networking Experiments And Technologies*, ser. CoNEXT '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 172–185. [Online]. Available: https://doi.org/10.1145/3359989.3365418

[26] "Wifirttscan app - apps on google play," https://play.google.com/store/apps/details?id=com.google.android.apps.location.rtt.wifirttscan.

[27] S. Irshad, E. Rozner, A. Bhartia, and B. Chen, "Rethinking wireless network management through sensor-driven contextual analysis," in *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*, ser. HotMobile '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 92–97. [Online]. Available: https://doi.org/10.1145/3376897.3377863

[28] L. T. Nguyen, Y. S. Kim, P. Tague, and J. Zhang, "Identitylink: User-device linking through visual and rf-signal cues," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '14. New York, NY, USA: ACM, 2014, pp. 529–539. [Online]. Available: http://doi.acm.org/10.1145/2632048.2636072

[29] S. Cao, H. Farrukh, and H. Wang, "Video: Enabling public cameras to talk to the public," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 546. [Online]. Available: https://doi.org/10.1145/3210240.3211118

[30] N. Guo, J. Luo, Z. Ling, M. Yang, W. Wu, and X. Fu, "Your clicks reveal your secrets: A novel user-device linking method through network and visual data," *Multimedia Tools Appl.*, vol. 78, no. 7, p. 8337–8362, Apr. 2019. [Online]. Available: https://doi.org/10.1007/s11042-018-6815-6

[31] J. Teng, B. Zhang, J. Zhu, X. Li, D. Xuan, and Y. F. Zheng, "Ev-loc: Integrating electronic and visual signals for accurate localization," *IEEE/ACM Transactions on Networking*, vol. 22, no. 4, pp. 1285–1296, 2014.

[32] H. W. Kuhn and B. Yaw, "The hungarian method for the assignment problem," *Naval Res. Logist. Quart*, pp. 83–97, 1955.

[33] F. Zafari, A. Gkelias, and K. Leung, "A survey of indoor localization systems and technologies," 2017. [Online]. Available: https://arxiv.org/abs/1709.01015

[34] J. Xu, H. Chen, K. Qian, E. Dong, M. Sun, C. Wu, L. Zhang, and Z. Yang, "Ivr: Integrated vision and radio localization with zero human effort," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, Sep. 2019. [Online]. Available: https://doi.org/10.1145/3351272

[35] X. Li, J. Teng, Q. Zhai, J. Zhu, D. Xuan, Y. F. Zheng, and W. Zhao, "Ev-human: Human localization via visual estimation of body electronic interference," in *2013 Proceedings IEEE INFOCOM*, 2013, pp. 500–504.

[36] A. Alahi, A. Haque, and L. Fei-Fei, "Rgb-w: When vision meets wireless," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3289–3297.

[37] J. P. Jeong, S. Yeon, T. Kim, H. Lee, S. M. Kim, and S.-C. Kim, "Sala: Smartphone-assisted localization algorithm for positioning indoor iot devices," *Wirel. Netw.*, vol. 24, no. 1, p. 27–47, Jan. 2018. [Online]. Available: https://doi.org/10.1007/s11276-016-1309-9

[38] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[40] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *CoRR*, vol. abs/1812.11941, 2018. [Online]. Available: http://arxiv.org/abs/1812.11941

[41] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2366–2374.

[42] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.