VILL: Toward Efficient and Automatic Visual Landmark Labeling

QUN NIU, Sun Yat-sen University and Guangdong Key Laboratory of Big Data Analysis and Processing
KUNXIN ZHU, Sun Yat-sen University
SUINING HE, The University of Connecticut
SHAOQI CEN, Sun Yat-sen University
S.-H. GARY CHAN, The Hong Kong University of Science and Technology
NING LIU, Sun Yat-sen University and Guangdong Province Key Laboratory of Information Security Technology

Of all indoor localization techniques, vision-based localization emerges as a promising one, mainly due to the ubiquity of rich visual features. Visual landmarks, which present distinguishing textures, play a fundamental role in visual indoor localization. However, few researches focus on visual landmark labeling. Preliminary arts usually designate a surveyor to select and record visual landmarks, which is tedious and time-consuming. Furthermore, due to structural changes (e.g., renovation), the visual landmark database may be outdated, leading to degraded localization accuracy.

To overcome these limitations, we propose *VILL*, a user-friendly, efficient, and accurate approach for visual landmark labeling. VILL asks a user to sweep the camera to take a video clip of his/her surroundings. In the construction stage, VILL identifies unlabeled visual landmarks from videos adaptively according to the graph-based visual correlation representation. Based on the spatial correlations with selected anchor landmarks, VILL estimates locations of unlabeled ones on the floorplan accurately. In the update stage, VILL formulates an alteration identification model based on the judgments from different users to identify altered landmarks accurately. Extensive experimental results in two different trial sites show that VILL reduces the site survey substantially (by at least 65.9%) and achieves comparable accuracy.

CCS Concepts: • **Networks** → **Location based services**; • **Human-centered computing**;

Additional Key Words and Phrases: Visual landmark identification, alteration identification

This work was supported in part by the National Natural Science Foundation of China under grants 62102459 and 61972433, Guangdong Basic and Applied Research Foundation under grant 2021A1515012242, and the Hong Kong General Research Fund under grant 16200120.

Authors' addresses: Q. Niu, School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, China, 519080 and Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, China, 510006; email: niuq3@mail.sysu.edu.cn; K. Zhu and S. Cen, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, 510006; emails: {zhukx3, censhq5}@mail2.sysu.edu.cn; S. He, Department of Computer Science and Engineering, The University of Connecticut, Storrs, CT 06269-4155; email: suininghe@uconn.edu; S.-H. G. Chan, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China; email: gchan@cse.ust.hk; N. Liu (corresponding author), School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, 510006, and Guangdong Province Key Laboratory of Information Security Technology, Guangzhou, China, 510006; email: liuning2@mail.sysu.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
1550-4859/2023/04-ART74 $15.00
https://doi.org/10.1145/3580497

ACM Reference format:
Qun Niu, Kunxin Zhu, Suining He, Shaoqi Cen, S.-H. Gary Chan, and Ning Liu. 2023. VILL: Toward Efficient and Automatic Visual Landmark Labeling. *ACM Trans. Sensor Netw.* 19, 4, Article 74 (April 2023), 25 pages.
https://doi.org/10.1145/3580497

1 INTRODUCTION

Visual landmark-based indoor localization has attracted much attention lately, mainly because visual landmarks are visually distinguishing and spatially pervasive, and do not require additional infrastructure support [34, 44]. Visual landmarks (e.g., store logos, signs, and wall paintings), which are visually distinguishing objects associated with specific locations, provide strong location clues as compared with other signals. Taking radio signals (e.g., Wi-Fi, Bluetooth) as an example, they can be affected by multi-path propagation, device orientation, and signal reflection, leading to degraded localization accuracy [5, 18, 24, 25]. Furthermore, they need to deploy a large number of wireless devices, which incurs additional deployment and maintenance cost.

Therefore, many researchers study accurate indoor localization with visual landmarks [11, 20, 26, 31]. The accuracy of visual indoor localization algorithms, however, is largely determined by the accuracy of the visual landmark database, as user locations are calculated based on the positions of visual landmarks in the database. Despite the promising applications, visual landmarks with accurate labels (their images and locations on the floorplan) are often unavailable or prohibitively costly to acquire [4, 8, 10, 27, 38]. Furthermore, visual landmarks may be altered due to structural changes caused by constant renovations, leading to degraded localization accuracy. To address this, surveyors have to survey the site regularly to identify altered ones and update them subsequently, leading to prohibitive cost of maintenance.

Although it is crucial to construct the visual landmark database accurately and efficiently, the research focusing on visual landmark labeling is limited. Preliminary arts [1, 16] select visual landmarks manually and estimate their locations based on the user trajectory and relative distances. Although they reduce the cost of location labeling, they incur constant calibration of noisy motion sensors [37]. Combined with the manual selection and labeling of *all* visual landmarks, they are sophisticated and error-prone in large trial sites.

Instead of the tedious landmark selection and image taking, we leverage the natural behavior of video taking: a user can either stand at a position and rotate arms to take a video of the environment or take the video while the user is walking. As videos taken by users cover different viewpoints of visual landmarks, we can identify them, estimate their locations, and update altered ones efficiently and automatically. However, it is difficult to label visual landmarks accurately, mainly due to the following challenges:

- *Landmark mis-identification due to texture diversity and unknown number*: Due to the diversified texture, color, and shape, existing visual landmark classification algorithms may regard unlabeled ones as the background. In addition, the number of visual landmarks is usually unknown, rendering it difficult to identify them accurately by clustering in a large number of unstructured videos.
- *Large location error with sensor noises*: Structure-from-Motion (SfM) [33, 41] models the spatial correlation between visual landmarks. Using the constructed 3-D model, we can infer the locations of unlabeled ones based on the spatial correlation with labeled visual landmarks. Due to statistical noises in videos, the 3-D model may be noisy, leading to large location errors of unlabeled landmarks.
- *Erroneous alteration identification with a single video*: Much work has shown that the confidence of object identification is high if the target is in the training dataset and identified

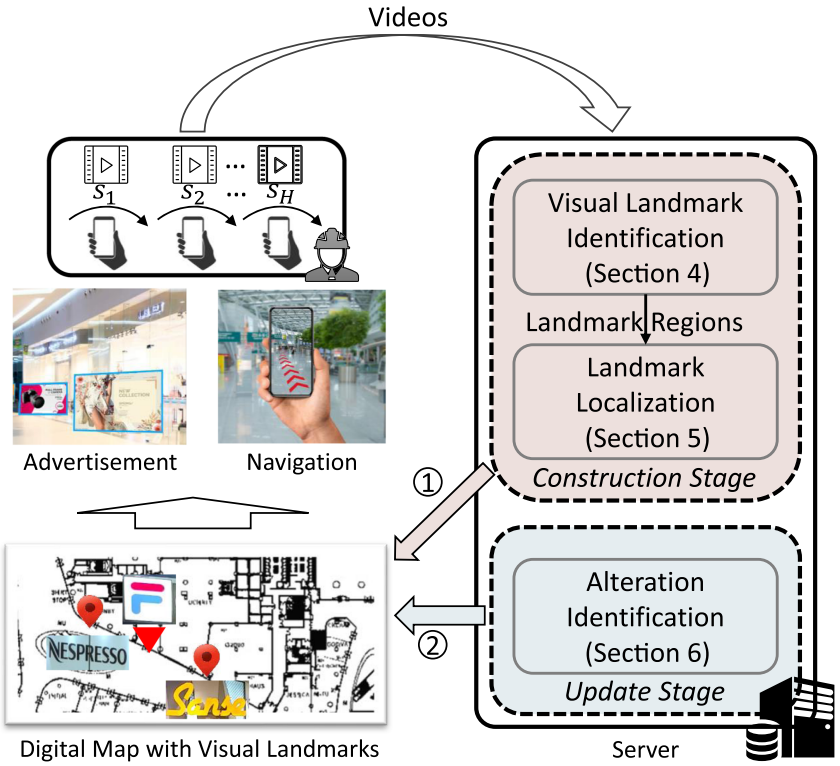


Fig. 1. System framework of VILL.

correctly. In the case of landmark alteration, the confidence values of altered visual landmarks are low, as they are not in the training dataset. Meanwhile, due to opportunistic noises (e.g., loss of focus, strong illumination), frames of unchanged landmarks in a single video could be blurry, leading to low confidence values as well. Therefore, it is challenging to identify altered landmarks accurately in a single video.

To address the preceding challenges, we propose an efficient and accurate *vi*sual *l*andmark *l*abeling approach by camera sweeping, termed *VILL*. We present the framework of the proposed VILL in Figure 1. VILL consists of two stages: a construction stage and an update stage. In the construction stage, a surveyor uploads videos $(s_1, s_2, \ldots, s_H)$ to a remote server. Then, it extracts candidate landmark regions and identifies unlabeled visual landmarks adaptively based on the graph-based similarity representation. Next, VILL uses anchor landmarks selected by the surveyor as guidance to estimate the locations of unlabeled ones (indicated by the red triangle). In the update stage, VILL leverages user-collected videos and identifies altered visual landmarks by taking multiple videos into consideration.

Our major contributions are as follows:

- *Adaptive graph-based visual landmark identification*: As a graph models the pairwise similarity of vertices and can be generalized to a large number of vertices, we use a vertex to represent a candidate region and connect two vertices if they are visually similar. If they belong to the same visual landmark, the weight value of the edge is large. Otherwise, the weight is significantly small or zero. Therefore, we propose to identify unlabeled visual

landmarks by finding *subgraphs*, which does not require prior knowledge about trial sites, landmark texture, or number, thus achieving adaptiveness.

- *Accurate landmark location labeling with anchor guidance*: We propose to select visual landmarks with the longest distance in a group as *anchor landmarks*. Combined with the spatial correlations inferred from the 3-D model,[1] we estimate the locations of unlabeled visual landmarks on the floorplan. Furthermore, we introduce the structural constraints into the location labeling of visual landmarks (landmarks are collocated on the wall) to reduce the impact of statistical noises in the 3-D model, thus achieving higher accuracy. We demonstrate its effectiveness theoretically.
- *Accurate alteration identification with multiple videos*: To identify altered visual landmarks accurately, we propose an alteration identification model, where we take the judgments from several users into consideration simultaneously. Specifically, we regard a single video as a *source* and *fuse* the judgments from different sources to determine if landmark is altered. Using videos collected by multiple users rather than a single one, we reduce the adverse impact of opportunistic noises (e.g., motion blur, strong illumination) and subsequently improve the identification accuracy of altered visual landmarks.

We have implemented VILL and conducted extensive experimental studies in two different trial sites: a crowded food court and a spacious shopping mall. Evaluation results show that our approach reduces the survey time (in terms of landmarks to be labeled) by at least 65.9% with comparable accuracy (in terms of landmark locations). Therefore, other visual sensing-based applications, such as indoor localization [28, 49], augmented reality [6], and visual assistance [17], can leverage our approach to reduce the survey cost and consequently enhance their applicability. Furthermore, VILL is natural to use, intuitive to novice users, and accurate in visual landmark labeling. Therefore, it can adopt crowdsourcing [19, 39, 48, 50] easily to further reduce the site survey of dedicated surveyors. Additionally, VILL can be deployed on mobile platforms (e.g., drones, wheeled robots) to construct and update the database efficiently.

The rest of this article is structured as follows. We review recent arts that are most related to ours in Section 2. Then, we overview the framework and major terminologies of VILL in Section 3. We elaborate the graph-based visual landmark identification algorithm in Section 4, followed by the anchor selection strategy in Section 5 and the alteration identification model in Section 6. We present illustrative experimental results in Section 7, followed by discussions and future directions in Section 8. We conclude in Section 9.

2 RELATED WORK

In this section, we review research that is most related to ours.

Visual Landmark Detection. Different from radio or magnetic signal-based landmark detection [7, 22, 32, 42, 47], visual landmark detection is usually challenging due to the high dimension of images and large variations (in terms of texture, color, and scale) of visual landmarks. Due to the significant progress in deep learning techniques, neural networks are used in a wide range of object detection tasks. Faster R-CNN [35], for example, extracts feature maps from input images with a feature pyramid network (FPN). Afterward, region proposal network (RPN) estimates coordinates of rectangular regions with objects in them, termed *object proposals*. Then, an alignment layer aligns extracted feature maps with coordinates without any quantization for better performance, generating local regions of features corresponding to potential objects. Besides the typical Faster R-CNN, recent research proposes to leverage the transformer architecture for accurate object de-

[1]We use "3-D point cloud" and "3-D model" interchangeably in this article.

tection. Swin Transformer [29] proposes a hierarchical transformer, where the representation is computed with shifted windows. Due to the hierarchical architecture, Swin Transformer is flexible to model various scales. Some research [38, 45] employs multi-class neural networks to detect visual landmarks. Although accurate, they do not generalize well to unseen ones due to domain differences [23, 40].

Visual Landmark Localization. Recent studies on visual landmark localization are broadly classified into two categories: trajectory-based approaches and indoor structure-based ones. The first category of methods recovers user traces and estimates locations of visual landmarks according to the current user position [1, 13]. However, they are prone to accumulative errors of motion sensors, thus requiring constant calibrations. Sextant [11] estimates the landmark location by triangulation with two known visual landmarks. Although efficient, the location errors of distant visual landmarks grow large due to accumulative errors of motion sensors, the location errors of known landmarks, and misalignment between the landmark center and the image center. Knitter [14] estimates the distance between a user and a visual landmark based on the facade geometry. To achieve sufficient accuracy, they ask the user to point the camera to the center of the shop facade, which is tedious for novice users.

Structure-based approaches, on the other hand, extract geometrical features (e.g., lines, corner points) to locate a visual landmark. ClickLoc [43] infers the translation of a visual landmark based on the accurate detection of corners between the entrance line and walls lines, which are not always available in the middle of a hallway. ViNav [9] infers the location based on their corresponding 3-D feature points in the point cloud.

Deep Learning Based Visual Inertial Odometry. With the development of deep learning techniques, many research works [12, 15, 21] study ego-motion estimation with visual input from target devices. GANVO [2] creates supervisory signals by warping view sequences and assigning reprojection minimization in the loss function to estimate the user trajectory using pure visual information. To enhance the accuracy, later approaches integrate inertial sensing in the trajectory estimation. Gao et al. [12] propose to eliminate sensor noises by learning both forward and backward inertial sequences. SelfVIO [3] introduces adversarial learning and self-adaptive sensor fusion to estimate the user trajectory. Jia et al. [21] devise a Gaussian estimator to predict the depth and uncertainty simultaneously. Although the preceding approaches achieve accuracy with trajectory recovery and depth estimation, they are different from ours in several ways. First, our target is to identify and label visual landmarks efficiently, whereas the preceding approaches focus on user trajectory estimation. Furthermore, we propose a crowd consensus-based approach to find altered visual landmarks and update them subsequently. However, these approaches do not consider landmark alteration.

3 WORKFLOW AND TERMINOLOGIES

Our VILL consists of two stages: a construction stage and an update stage. In the construction stage, a surveyor divides visual landmarks into several groups according to their locations on the digital map—that is, visual landmarks that are spatially close on a wall segment are divided into a group. Then, the surveyor selects two anchor landmarks for each group and annotates their positions. Afterward, surveyors or volunteers stand at a position and rotate arms to take video clips. They can also hold the camera and record videos as they walk. The client application sends videos to a server. Upon receiving videos, VILL builds a 3-D model with SfM for each group of visual landmarks. Then, it detects candidate landmark regions with a two-class classifier. Using these regions, it builds a *connectivity graph* and identifies candidate visual landmarks by finding connected subgraphs (Section 4). Based on the relative distances from anchor landmarks in the 3-D model, VILL estimates physical locations of unlabeled ones on the floorplan (Section 5).

Table 1. Major Notations in VILL

Notation	Definition
$\mathbf{x}$	2-D location on the floorplan
H	Number of source videos
L	Number of visual landmarks
k_{ij}	Pairwise conflict between source i and j
γ	Overall conflict of all sources
Δ	Confidence values of all visual landmarks
I, N, C, K	Hypotheses of the alteration identification model
m	Basic probability assignment
r	Fusion probability

In the update stage, VILL can either work in stand-alone mode or be integrated into existing visual crowdsourcing applications. It analyzes uploaded videos by a number of users and detects visual landmarks in each of them. Based on the video noises and the user consensus, VILL identifies altered visual landmarks, labels them, and updates their locations accordingly (Section 6). We present major terminologies as follows.

Definition 1 (Anchor landmarks). These are reference landmarks selected and labeled by the surveyor (with images and positions), which are used as anchors to guide the mapping of visual landmarks from the point cloud to the floorplan.

Definition 2 (User consensus). It is the indicator that differnet users reach an agreement whether a landmark is altered.

We present major notations used in this article in Table 1.

4 GRAPH-BASED VISUAL LANDMARK IDENTIFICATION

Due to the large diversity and unknown number of visual landmarks, we propose an adaptive graph-based algorithm to identify unlabeled visual landmarks from a large number of unstructured candidate regions. We present the two-class visual landmark detector that detects candidate landmark regions in Section 4.1. Then, we elaborate our graph-based visual landmark identification algorithm in Section 4.2.

4.1 Detection of Candidate Landmark Regions

Besides landmark detection, object detection networks (e.g., Faster R-CNN) classify detected objects simultaneously by learning fine-grained visual features. Although accurate, they are better with *labeled* landmarks. As for unlabeled ones, they may be regarded as the background incorrectly due to domain differences, rendering it difficult to detect unlabeled visual landmarks.

As indoor landmarks usually share common visual and spatial characteristics (e.g., text, rich texture, horizontally, or vertically aligned), it is possible to train a two-class network that detects candidate landmark regions *without* identifying their categories. This reduces the need to learn fine-grained features of visual landmarks, thus increasing the generality of networks to unlabeled ones. Motivated by this, we propose to train a two-class network to detect candidate landmark regions in images without classifying them. Our region detection network is based on the state-of-the-art Faster R-CNN, where we modify the last few convolutional layers so that our network outputs regions with *two* labels: landmark and background.

Fig. 2. SIFT feature points.

Fig. 3. Comparison of SIFT features in two images.

4.2 Graph-Based Similarity Representation and Landmark Identification

Scale-Invariant Feature Transform (SIFT) [30] is an algorithm that extracts local features from input images. SIFT features are invariant to image scale and rotation, and are robust to a substantial range of affine distortions and viewpoint changes. Consequently, we extract SIFT descriptors from images and compare them afterward to evaluate the similarity between two images. We illustrate extracted SIFT points in Figure 2 and connect matched features with solid lines in Figure 3. Based on the number of matched features, we can determine the similarity between two images. We use the SfM technique to build a 3-D model of the environment using SIFT features and descriptors. Based on the pairwise visual similarities (in terms of matched features) inferred from the 3-D point cloud, we propose a graph-based visual landmark identification algorithm.

Formally, given a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ formed by candidate regions, where $\mathbf{V}$ denotes a set of vertices (each vertex indicates a candidate landmark region) and $\mathbf{E}$ denotes a set of edges (each edge indicates similarity between two regions). In our algorithm, we use the number of matched visual features to indicate the similarity between two regions (i.e., weight of the edge). As regions of the same landmark are visually similar, they usually have a large number of matched features (i.e., large weight value), whereas regions of different visual landmarks have few matched features (i.e., no edges). Consequently, the vertices belonging to the same landmark form a *connected subgraph*. Our goal is to find connected subgraphs from $\mathbf{G}$, where each one indicates a visual landmark.

ALGORITHM 1: Graph-Based Visual Landmark Identification

Input: $\mathbf{G}$
Output: A set of root vertices for all subgraphs Ω, A set of visited vertices Υ
Initialization : $\Omega = \emptyset$, $\Upsilon = \emptyset$
1: **while** $|\Upsilon| \neq |\mathbf{V}|$ **do**
2: From the unvisited set, select a vertex v' connected by the edge with the largest weight
3: Add v' to Ω and Υ
4: Find connected vertices of v' via breadth-first traversal
5: Add connected vertices to Υ
6: Mark all vertices in Υ as visited, record the subgraph
7: **end while**
8: **return** Ω

We present our adaptive landmark identification algorithm in Algorithm 1. Initially, we mark all vertices as unvisited. We remove edges with weight values smaller than threshold. Afterward, we randomly select a vertex from the graph that has not been visited before. Then, we find its connected vertices using the breadth-first algorithm, as they indicate similar regions. Then, we add them to Υ and mark them as visited. Meanwhile, we mark the selected v' as the root vertex for this subgraph (a visual landmark). We continue the procedure until all vertices are added to Υ. Finally, we get a set of root vertices Ω and select subgraphs with vertex larger than threshold as unlabeled visual landmarks.

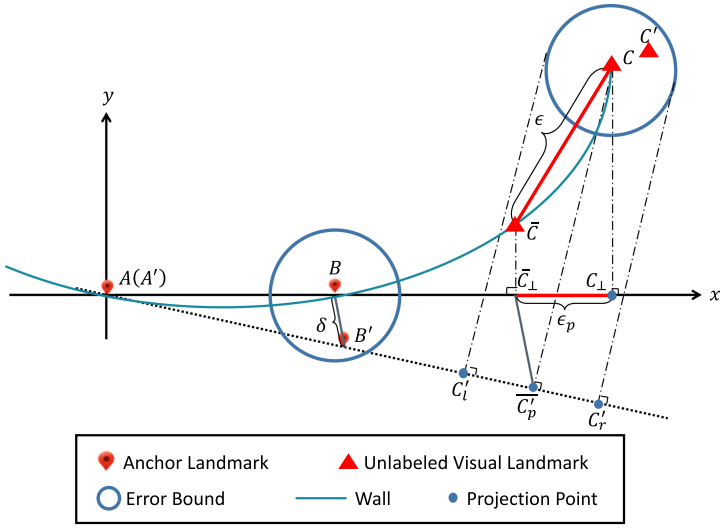


Fig. 4. Anchor-based visual landmark location estimation.

5 ANCHOR-GUIDED ACCURATE LANDMARK LOCALIZATION

Due to statistical noises in images, the spatial distances between visual landmarks in the 3-D model may be noisy, leading to large location errors of unlabeled visual landmarks. The selection of anchor landmarks has a significant impact on the localization error. In light of this, we propose an anchor selection strategy to reduce the overall localization error of unlabeled visual landmarks. We propose the anchor selection strategy in Section 5.1, followed by the theoretical justification of our anchor selection strategy in Section 5.2.

5.1 Anchor Selection and Visual Landmark Positioning

We illustrate visual landmark positioning given a noisy 3-D model in Figure 4. Based on the analysis of location error, we present our anchor selection strategy accordingly, where we select visual landmarks with the longest distance as anchor landmarks for each group.

Given three visual landmarks, A, B, and C, whose physical coordinates are $\mathbf{x}_A$, $\mathbf{x}_B$, and $\mathbf{x}_C$, respectively. Of these landmarks, A and B are anchor landmarks and their coordinates on the digital map are annotated manually, and C is the unlabeled visual landmark to be localized. The corresponding landmarks in the point cloud are denoted by A', B', and C', respectively. Their coordinates in the 3-D point cloud are denoted by $\mathbf{x}_{A'}$, $\mathbf{x}_{B'}$, and $\mathbf{x}_{C'}$, respectively. Our objective is to estimate the location of visual landmark C accurately.

Due to statistical noises in images and calibration errors of cameras, the estimated locations of landmarks in the point cloud are noisy. We illustrate this with B. Suppose its estimated location in the point cloud is at B', which is within a round region centered at B with error bound δ ($0 \leq \delta \leq \Delta$). Similarly, the estimated location of the unlabeled landmark C in the point cloud, denoted by C', is also noisy. For simplicity, we adjust the point cloud so that the position of landmark A in point cloud A' overlaps with its physical location. Without loss of generality, we elaborate the localization of C, which is not collinear with anchor landmarks A and B. The wall segment is defined as

$$f(\mathbf{x}) = a_0 + a_1\mathbf{x} + a_2\mathbf{x}^2 + \cdots + a_n\mathbf{x}^n. \tag{1}$$

First, we map the C' to the dotted line that connects A' and B' in the point cloud. Since C' could be randomly distributed in a circle centered at C, the expected projection of C' onto the line determined by A' and B' is $\bar{C'_p}$. Then, we draw a line segment $\bar{C'_p}\bar{C}_\perp$, which is parallel to BB' and intersects with AB at $\bar{C}_\perp$. As the 3-D point cloud models the relative spatial relations between visual landmarks on the digital map, we have $\frac{|AB|}{|A\bar{C}_\perp|} = \frac{|A'B'|}{|A'\bar{C'}_p|}$.

$C_\perp$ is the projection of C on the line that connects AB. Meanwhile, $CC_\perp$ is perpendicular to AB and intersects with AB at $C_\perp$. As C lies on the same wall segment as A and B, we draw a line that is perpendicular to AB and intersects with it at $\bar{C}_\perp$. This line intersects with AC at $\bar{C}$, which is the estimated location of C on the floorplan. The location error of visual landmark C is $\epsilon = |\bar{C}C|$, and the error projected on AB is ϵ_p. As we can infer from Figure 4, ϵ_p is closely related to the distances between A and B. In our experiment, we select visual landmarks with the longest distance as anchor landmarks to achieve sufficient accuracy for unlabeled ones.

5.2 Theoretical Justification of Anchor Selection Strategy

Suppose the physical coordinates of anchor landmarks annotated by surveyors are accurate, but the estimated locations of visual landmarks in the point cloud are noisy. Note that it is possible to transform the coordinate system of visual landmarks by translation. Without loss of generality, we transform the coordinate system and justify our strategy where the coordinate of A' is accurate, whereas that of B' is noisy.

For simplicity, we set A as the origin of our coordinate system. We have B on the positive side of the x-axis. As illustrated in Figure 4, $\alpha = \angle CAC_\perp < \angle CBC_\perp$. As C (x_C, y_C) lies in the first quadrant, we have $x_C > 0, y_C > 0$. Due to statistical noises in images, the estimated location B' in the point cloud and B do not overlap. Let AB be the x-axis; we have A (0,0), B $(x_B, 0)$, B' $(x_{B'}, y_{B'})$, and C(x_C, y_C). Given $C\bar{C'}_p \perp A'\bar{C'}_p$ and $\bar{C}_\perp\bar{C'}_p \parallel BB'$, we have

$$\bar{C'}_p = \left(\frac{x_{B'}^2 x_C + y_{B'}x_{B'}y_C}{x_{B'}^2 + y_{B'}^2}, \frac{x_{B'}y_{B'}x_C + y_{B'}^2 y_C}{x_{B'}^2 + y_{B'}^2}\right), \tag{2}$$

$$\frac{A\bar{C}_\perp}{AB} = \frac{A'\bar{C'}_p}{A'B'}, \tag{3}$$

$$\epsilon_p = |\bar{C}_\perp C_\perp| = ||AC_\perp| - |A\bar{C}_\perp||. \tag{4}$$

Let $u_C = \frac{y_C}{x_C}(u_C \geq 0)$; we have

$$\epsilon_p = \left|\frac{|x_{B'} + u_C y_{B'}|}{x_{B'}^2 + y_{B'}^2} - \frac{1}{x_B}\right|\left|\frac{x_B}{x_C}\right|. \tag{5}$$

We evaluate the impact of $|x_{B'} + u_C y_{B'}|$ on the location error of visual landmark C as follows:

- $x_{B'} + u_C y_{B'} \leq 0$: Since B' $(x_{B'}, y_{B'})$ lies in the fourth quadrant, we can infer that $u_C \geq -\frac{x_{B'}}{y_{B'}} > \frac{x_{B'}}{\Delta}$. As the error of 3-D model is marginal ($x_{B'} \gg \Delta$), u_C could be significantly large. This indicates that C is far away from AB, which contradicts our assumption that anchor landmarks and unlabeled ones in a group are spatially close. Therefore, we discard this scenario in our justification.
- $x_{B'} + u_C y_{B'} > 0$: In this case, we can infer that $\frac{y_{B'}}{x_{B'}} \geq -\frac{1}{u_C}$. Consequently, we have

$$\epsilon_p = \left|\frac{x_{B'} + u_C y_{B'}}{x_{B'}^2 + y_{B'}^2} - \frac{1}{x_B}\right|\left|\frac{x_B}{x_C}\right|. \tag{6}$$

We further evaluate the remaining terms as follows. If $\frac{x_{B'}+u_C y_{B'}}{x_{B'}^2+y_{B'}^2} - \frac{1}{x_B} > 0$, we have

$$\begin{aligned}
\epsilon_p &= \left(\frac{x_{B'}+u_C y_{B'}}{x_{B'}^2+y_{B'}^2} - \frac{1}{x_B}\right)\left|\frac{x_B}{x_C}\right| \\
&\leq \left(\frac{x_B+\Delta+u_C\Delta}{(x_B-\Delta)^2} - \frac{1}{x_B}\right)\left|\frac{x_B}{x_C}\right| \\
&= \frac{(u_C+3)x_B\Delta-\Delta^2}{(x_B-\Delta)^2 x_B}\left|\frac{x_B}{x_C}\right| \\
&< \frac{(u_C+3)x_B\Delta-\Delta^2/x_B}{(x_B-\Delta)^2}\left|\frac{x_B}{x_C}\right| \\
&= \frac{(u_C+3)\Delta}{(x_B-2\Delta+\Delta^2/x_B)x_C}.
\end{aligned} \tag{7}$$

As x_B grows larger, $\frac{(u_C+3)\Delta}{(x_B-2\Delta+\Delta^2/x_B)x_C}$ becomes smaller and closer to zero. Consequently, ϵ_p gets closer to zero as x_B grows larger, which indicates longer distances from A. In this case, we conclude that as the distance between two anchor landmarks grows larger, the upper bound of location error in the same region becomes smaller.

Similarly, in the case where $\frac{x_{B'}+u_C y_{B'}}{x_{B'}^2+y_{B'}^2} - \frac{1}{x_B} < 0$, we have

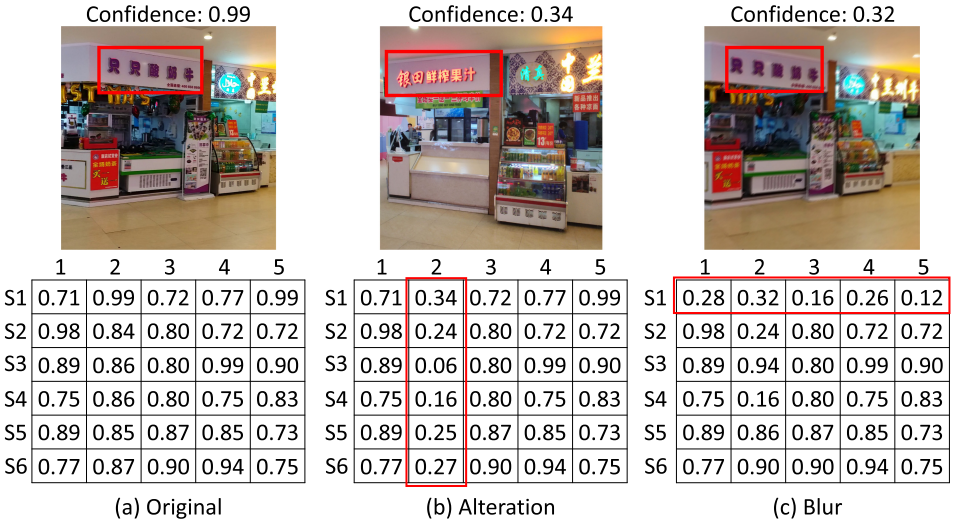
$$\begin{aligned}
\epsilon_p &= \left(\frac{1}{x_B} - \frac{x_{B'}+u_C y_{B'}}{x_{B'}^2+y_{B'}^2}\right)\left|\frac{x_B}{x_C}\right| \\
&< \left(\frac{1}{x_B} - \frac{x_B-\Delta-u_C\Delta}{(x_B+\Delta)^2}\right)\left|\frac{x_B}{x_C}\right| \\
&= \frac{x_B\Delta(u_C+3)+\Delta^2}{x_B(x_B+\Delta)^2}\left|\frac{x_B}{x_C}\right| \\
&< \frac{\Delta(u_C+3)+1}{(x_B+\Delta)^2}\left|\frac{x_B}{x_C}\right| \\
&= \frac{\Delta(u_C+3)+1}{(x_B+2\Delta+\Delta^2/x_B)x_C}.
\end{aligned} \tag{8}$$

Similarly, we can infer that as x_B becomes larger, $\frac{\Delta(u_C+3)+1}{(x_B+2\Delta+\Delta^2/x_B)x_C}$ grows smaller. Consequently, we select the pair of visual landmarks with the longest distances in a group as anchor landmarks to reduce the upper bound of location errors in unlabeled visual landmarks.

Although we justify the anchor selection strategy where unlabeled visual landmarks are not collinear with others, our strategy is easily applicable to trial sites where visual landmarks are on flat surface (i.e., linear). This is because we calculate ϵ by dividing ϵ_p by $\cos\alpha$, which is a constant value. Therefore, our strategy is general to both scenarios with linear or non-linear wall partitions.

6 USER CONSENSUS-GUIDED ALTERATION IDENTIFICATION

Due to statistical noises in images, the confidence value of an unchanged visual landmark may decrease significantly. In this case, it is difficult to determine if the visual landmark is altered based on the confidence value from a video. To address this, we propose an identification algorithm that jointly considers multiple video sources. We first motivate our design with confidence values in typical scenarios in Section 6.1. Then, we present preliminaries of the **Dempster-Shafer (DS)** theory in Section 6.2, followed by the proposed alteration identification model in Section 6.3. Based



	1	2	3	4	5
S1	0.71	0.99	0.72	0.77	0.99
S2	0.98	0.84	0.80	0.72	0.72
S3	0.89	0.86	0.80	0.99	0.90
S4	0.75	0.86	0.80	0.75	0.83
S5	0.89	0.85	0.87	0.85	0.73
S6	0.77	0.87	0.90	0.94	0.75

(a) Original

	1	2	3	4	5
S1	0.71	0.34	0.72	0.77	0.99
S2	0.98	0.24	0.80	0.72	0.72
S3	0.89	0.06	0.80	0.99	0.90
S4	0.75	0.16	0.80	0.75	0.83
S5	0.89	0.25	0.87	0.85	0.73
S6	0.77	0.27	0.90	0.94	0.75

(b) Alteration

	1	2	3	4	5
S1	0.28	0.32	0.16	0.26	0.12
S2	0.98	0.24	0.80	0.72	0.72
S3	0.89	0.94	0.80	0.99	0.90
S4	0.75	0.16	0.80	0.75	0.83
S5	0.89	0.86	0.87	0.85	0.73
S6	0.77	0.90	0.90	0.94	0.75

(c) Blur

Fig. 5. Confidence values of visual landmarks in different scenarios.

on the modeling of alteration identification, we provide the identification criteria of altered visual landmarks and computational analysis in Section 6.4.

6.1 Evaluation of Confidence Values

Many state-of-the-art object detection networks detect visual landmarks and present confidence values. If they are large, neural networks are confident that the landmark detection is correct and it is not altered. Otherwise, the confidence values are small. Theoretically, we can infer visual landmark alteration based on the confidence value. However, the confidence value may be low due to environmental noises (e.g., strong illumination, temporary occlusion) and sensor errors (e.g., loss of focus). We illustrate this with images collected in a food court in Figure 5. In this case, it is difficult to determine if a visual landmark is altered based on confidence values inferred from a single video source.

To address this, we propose to determine the alteration based on information inferred from multiple videos generated by different users rather than a single one to reduce the adverse impact of opportunistic errors. As an example, we present the confidence matrix of five visual landmarks collected in six different videos (termed *sources*) in Figure 5. For simplicity, we use "S1," "S2," ..., "S6" to indicate six video sources in these tables. We use numerical values 1 to 5 to indicate five different visual landmarks.

More specifically, we present the distribution of confidence values in two typical cases:

- *Small confidence values of a visual landmark in various sources*: This indicates that the observed confidence values of a single visual landmark are low across different videos (highlighted in the red bounding box in Figure 5(b)). As the number of cells with low confidence values increases, we can infer that an alteration may happen, resulting in low confidence values across different sources.
- *Small confidence values across different visual landmarks*: This indicates that the observed confidence values of all visual landmarks in a single video source are low (highlighted in the red bounding box in Figure 5(c)). The number of cells with low confidence values is large. We can infer that the video is prone to temporary environmental or sensor noises. Consequently, the confidence values of all visual landmarks are low.

6.2 Preliminaries of DS Theory

DS theory is a generalization of the Bayesian theory of subjective probability [36, 46]. It combines evidence from multiple sources and finally gives a degree of confidence for hypotheses. The mathematical model is defined as follows. Let Θ be a finite set of mutually exclusive and exhaustive hypotheses about some problem domain. A **Basic Probability Assignment (BPA)** is defined as a function m from 2^{Θ} to $[0, 1]$, which satisfies

$$m(\emptyset) = 0, \sum_{A \subseteq \Theta} m(A) = 1. \tag{9}$$

Given a finite number of functions $(m_1, m_2, \ldots, m_T)$, for each hypothesis A $(A \subseteq \Theta)$, Dempster's rule of combination is defined as follows:

$$r(A) = \frac{1}{1-\gamma} \sum_{A_1 \cap A_1 \cap \cdots \cap A_T = A} m_1(A_1) m_2(A_2) \cdots m_T(A_T), \tag{10}$$

where γ is a measure of overall conflict among sources. Specifically, it is a scalar value defined as follows:

$$\gamma = \sum_{A_1 \cap A_1 \cap \cdots \cap A_T \neq \emptyset} m_1(A_1) m_2(A_2) \cdots m_T(A_T). \tag{11}$$

However, if the conflict among sources is significant, γ may be close to 1. In this case, existing DS theory may suffer from the divide-by-zero error, leading to erroneous judgement, termed the *paradox problem*. Based on our detailed analysis of video sources and noise factors, we refine the DS model specifically based on our application scenarios.

6.3 Alteration Identification Modeling

We model our alteration problem as follows to determine if a visual landmark is altered. Given a visual landmark, we define four hypotheses regarding its possibility of being altered by taking environmental and sensor noises into consideration. These hypotheses are defined as follows:

Definition 3 (Hypothesis I). The landmark detection network can classify the visual landmark as the original one with probability $m(I)$.

Definition 4 (Hypothesis N). The landmark detection network can identify that the visual landmark is altered with probability $m(N)$.

Definition 5 (Hypothesis C). The impact of noises is so significant that the landmark detection network cannot determine if this is the original one with probability $m(C)$.

Definition 6 (Hypothesis K). The judgments from multiple sources are too contradictory to determine the event with probability $m(K)$.

Based on the preceding definitions, we formally define the hypothesis set as $\Theta = \{I, N, C, K\}$. To address the paradox problem, we propose to introduce additional measurements for robust alteration identification. Formally, we remove the denominator in Equation (10) and introduce the conflict indicator q to reduce the impact of noises with a few videos. More specifically, for each source (video), we have four BPAs, denoted by $m(I)$, $m(N)$, $m(C)$, and $m(K)$, respectively, and they add up to 1.

As the DS algorithm fuses multiple BPAs, we have the fusion probability for each hypothesis, where the corresponding event probabilities are denoted by $r(I)$, $r(N)$, $r(C)$, and $r(K)$, respectively. Consequently, we have

$$r(I) + r(N) + r(C) + r(K) = 1, \tag{12}$$

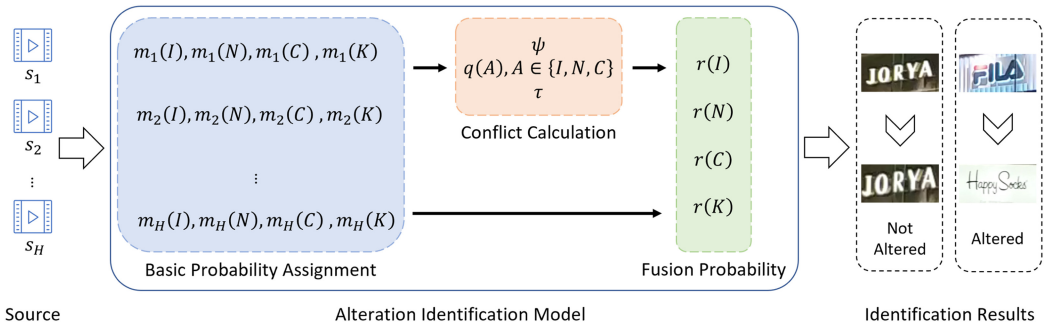


Fig. 6. Overall flow of alteration identification.

where $m(K), r(K) \in [0, 1]$ are indicators of the contradiction between multiple sources. As a single source does not contradict itself, we set the initial value of $m(K)$ to 0. $r(K)$, however, is not zero, as it measures the degree of contradiction among multiple sources. As the level of contradiction increases, the value of $r(K)$ increases. With fewer sources (say, two), the $r(K)$ could be large due to contradictory hypotheses. As the number of user-generated videos increases, $r(K)$ could become small as they achieve consensus. We overview the alteration identification in Figure 6. The proposed alteration identification consists of three major components: basic probability calculation, conflict probability calculation, and probability fusion. We detail each component as follows.

Suppose we have H sources (videos) and L visual landmarks in a region. Using the landmark detection network, we get a landmark sequence in each video. The confidence values for all visual landmarks are denoted by $\Lambda = [\lambda_1, \lambda_2, \ldots, \lambda_L]$. Then, we calculate the BPAs of hypotheses corresponding to l $(1 \le l \le L)$ based on source i $(1 \le i \le H)$, as follows:

$$m_i(C) = 1 - \sum_{l=1}^{L} \lambda_l / L, \tag{13}$$

$$m_i(I) = \lambda_l * (1 - m_i(C)), \tag{14}$$

$$m_i(N) = (1 - \lambda_l) * (1 - m_i(C)). \tag{15}$$

Then, we define the conflict between source i and j as follows:

$$\begin{aligned} k_{ij} = m_i(I) * (1 - m_j(I)) + m_i(N) * (1 - m_j(N)) \\ + m_i(C) * (1 - m_j(C)). \end{aligned} \tag{16}$$

Subsequently, the average pairwise conflict κ of all sources is defined as

$$\kappa = \frac{1}{H(H-1)/2} \sum_{i=1}^{H} \sum_{j=i+1}^{H} k_{ij}. \tag{17}$$

Additionally, we define a pairwise conflict indicator $\psi = e^{-\kappa}$, where ψ decreases as κ increases. If the pairwise conflict indicator is small, ψ is large, indicating high degree of belief.

We define w as the overall dot product of all sources:

$$w(A) = \prod_{i=1}^{H} m_i(A), A \in \{I, N, C\}, \tag{18}$$

and $w(K)$ is zero.

We use $q(I)$, $q(N)$ and $q(C)$ to denote the average degree of conflict among sources:

$$q(A) = \frac{1}{H}\sum_{i=1}^{H} m_i(A), A \in \{I, N, C\}, \tag{19}$$

and $q(K)$ is zero.

The overall conflict indicator is defined as follows:

$$\tau = 1 - w(I) - w(N) - w(C). \tag{20}$$

Please note that κ and τ are different conflict indicators. κ measures the pairwise conflict between each pair of sources. Consequently, it is large if assignments are contradictory among sources. In contrast to κ, τ pays more attention to the overall conflict of all sources. Say, if some sources conflict with others, τ becomes large even when the conflict indicator of other sources is small.

Finally, we give the fusion probability:

$$r(A) = w(A) + \tau * \psi * q(A), A \in \{I, N, C\}, \tag{21}$$

$$r(K) = w(K) + \tau * \psi * q(K) + \tau * (1 - \psi), \tag{22}$$

where $r(I)$ and $r(N)$ denote the fusion probability of being the original visual landmark and being altered, respectively. $r(C)$ and $r(K)$ are indicators of video noise and user consensus, respectively.

6.4 Criteria of Alteration Identification

Our alteration identification criteria are defined as follows. If the overall video noise indicator $r(C)$ and contradiction probability $r(K)$ are small, we can identify whether they are altered based on the comparison of $r(I)$ and $r(N)$. This is because the sources are less prone to noise (low $r(C)$) and consensus is achieved among them (low $r(K)$). If $r(I) > r(N)$, we determine that the landmark is altered. Otherwise, it is not altered. To reduce the impact of environmental noises (e.g., strong illumination) and user operations (e.g., motion blur) on a single source, we wait for more video sources (as illustrated in the experimental results) to determine if a visual landmark is altered. If, however, $r(C)$ is large (indicating that sources are noisy), we cannot know whether the visual landmark is altered. Additionally, if $r(K)$ is large, indicating hypotheses from different sources are highly contradictory. In both cases, surveyors need to determine if the visual landmark is altered.

Finally, we evaluate the computational complexity of the proposed algorithm. In our alteration identification algorithm, the calculation of k_{ij} is more expensive than other probabilities. As k_{ij} measures the pairwise conflict between two sources, the overall computational complexity is $O(H^2)$, where H denotes the number of sources.

7 ILLUSTRATIVE EXPERIMENTAL RESULTS

We have implemented VILL and conducted extensive experiments in two different trial sites: a crowded food court and a spacious shopping mall (Figure 7). We present our experimental settings in Section 7.1, followed by the illustrative results of visual landmark identification and alteration identification in Section 7.2. We present the location error of visual landmarks and users in Section 7.3 and system overhead in Section 7.4.

7.1 Experimental Settings and Comparison Schemes

As an example, we select store logos as visual landmarks. This is because they are relatively pervasive, distinguishing, and stable in these trial sites. Additionally, we have conducted evaluations in regions with posters in the food court. Other types of visual landmarks, such as sculptures and paintings, could also be introduced into our landmark set, as they are visually distinguishing from

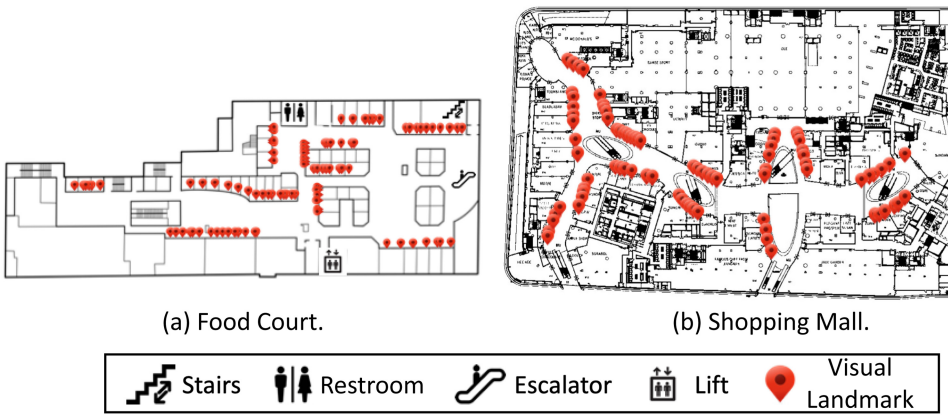


Fig. 7. Floorplans of our trial sites.

the background and can also be easily detected by state-of-the-art object detection and classification neural networks. To sum up, we have 92 and 82 visual landmarks in the food court (around 6,000 m^2) and the shopping mall (15,000 m^2), respectively.

We show the distances between adjacent visual landmarks in Figure 8. It shows that overall distances in the food court are shorter than those in the shopping mall. This is because the food court is more compact, where visual landmarks are closer to each other. In this case, one can take shorter videos in this trial site in the construction stage. As the shopping mall is more spacious, the distances between adjacent visual landmarks are longer.

We elaborate the group paradigm of visual landmarks as follows. We first divide visual landmarks into several groups based on their physical locations and the floorplan of the trial site. We divide neighboring visual landmarks into a group if they are spatially close—that is, located on the same wall segment. Consequently, we generate an accurate 3-D model to cover them for location inference. However, there are exits surrounded by white walls. Due to a lack of rich visual features, the 3-D model can be noisy, leading to erroneous location estimation. In this case, we divide visual landmarks into different groups if they are separated by exits. In the shopping mall with a large spacious atrium, we divide the landmarks on two sides into different groups so that we can build a more accurate 3-D model of the environment. Then, we select visual landmarks at either end of the wall as anchor landmarks.

In the construction stage (January 2019), there are 26 and 28 anchor landmarks in the food court and the shopping mall, respectively. Afterward, we invite two volunteers to take part in the survey. Each of them takes two smartphones to collect videos for visual landmark labeling. We collect 192 and 353 video clips in the construction stage. VILL uses these videos to label 66 visual landmarks in the food court and 54 visual landmarks in the shopping mall.

In the update stage (January 2020), another two volunteers take part in the video collection to detect altered visual landmarks. There are three and five landmark alterations in the food court and the shopping mall, respectively. We take 109 and 65 video clips in the food court and the shopping mall, respectively, with four different smartphones: Huawei Mate 9, Xiaomi 9 Pro, Xiaomi 6, and Samsung C5 Pro. In July 2021, we take another 24 videos in the shopping mall, where nine landmarks are altered. The frame rate of the video is around 30 frames per second.

We present the number of landmarks in user-collected videos in Figure 9. It shows that the distribution of the landmark number varies with videos and trial sites. In the food court, the number of visual landmarks in many videos is relatively small (e.g., three and four), as video clips collected

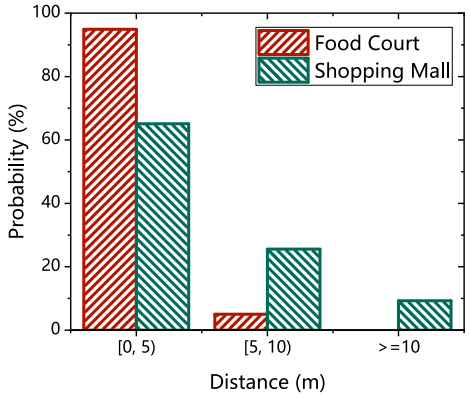


Fig. 8. Distances between adjacent landmarks.

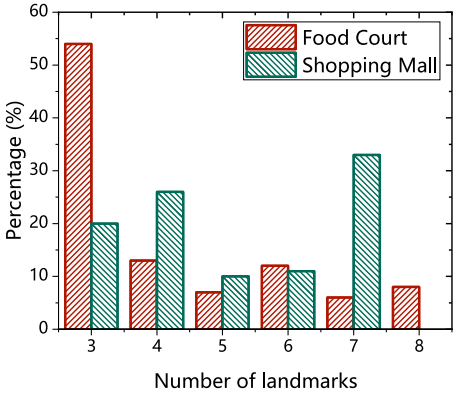


Fig. 9. Number of landmarks in videos.

in this trial site are relatively short. In contrast, the video clips are long in the shopping mall. Consequently, the number of visual landmarks in a video is relatively large.

We select Faster R-CNN to detect visual landmarks, as it is relatively accurate and efficient. The backbone of Faster R-CNN is VGG16. In our experiment, we use images of anchor landmarks as the training dataset for visual landmark detection. As store logos share visual characteristics, we introduce readily-available images of landmarks from other trial sites to enhance the training dataset. For example, in the food court, we use images of landmarks in the shopping mall and anchor landmarks as our training dataset for visual landmark detection. In the shopping mall, we use images of landmarks in the food court as well as images of anchor landmarks in the trial site as training images for visual landmark detection. We use the model provided by PyTorch[2] and modify the output of classification layer to the number to $(K + 1)$, where K is the number of visual landmarks and 1 stands for the background class. We also modify the nodes in the bounding box regression layer to $4 \times (K + 1)$. To fine-tune a Faster R-CNN network that can identify a candidate landmark region, we fine-tune the network with collected video frames. Specifically, we have 1,417 and 1,055 training images in the food court and the shopping mall, respectively.

We use VisualSFM[3] to construct the 3-D point cloud using video frames for each group of visual landmarks. VILL extracts SIFT features, conducts pairwise feature comparison, and builds a graph for candidate landmark regions where the weight between two vertices is the number of matched SIFT features. To reduce the impact of temporal noises, we set the threshold value as the mean number of matched features and remove those edges with weight values smaller than the threshold.

We evaluate the performance of visual landmark labeling with the following metrics:

- *Precision of visual landmark identification*: Given P connected subgraphs, each subgraph consists of $s_p (1 \leq p \leq P)$ images, where $\hat{s}_p$ of them belong to this visual landmark and others do not. Then, the precision in the construction stage is defined as $\frac{1}{P} \sum_{p=1}^{P} \frac{\hat{s}_p}{s_p}$.
- *Alteration identification accuracy*: Suppose we have Z trials for landmark alteration, and VILL gives z correct estimations. The accuracy of alteration identification is defined as $\frac{z}{Z}$.
- *Location error of visual landmarks*: Let M be the number of visual landmarks, and $\mathbf{x}_m$ and $\hat{\mathbf{x}}_m$ be the ground truth and estimated location of trial m, respectively; we define the location

[2] https://pytorch.org/.
[3] http://ccwu.me/vsfm/.

error as $\frac{1}{M}\sum_{m=1}^{M}||\mathbf{x}_m - \hat{\mathbf{x}}_m||_2$, where $||\cdot||_2$ indicates the Euclidean distance between two locations.

To evaluate the impact of source number on the alteration identification accuracy, we conduct simulations by selecting different numbers of sources (one, three, five, seven, and nine) randomly. We compare our VILL with the *Threshold*, which selects a source randomly and determines if a landmark is altered by comparing the confidence value with a threshold value.

We compare our approach with the state-of-the-art Sextant [11] to evaluate the location error of unlabeled visual landmarks. For fair comparison, we use the same number of anchor landmarks for both approaches. Additionally, we design a baseline approach without wall constraints (VILL w/o) to evaluate the location error. To evaluate the impact of video resolution on the localization accuracy, we downsample video frames and repeat the trials with different resolutions. We also conduct another simulation in the food court, where we select different pairs of visual landmarks as anchor landmarks and evaluate the impact of their distances on the location error of unlabeled ones.

To evaluate the impact of video number on the success rate and location error of unlabeled visual landmarks, we select different numbers of videos randomly and repeat the experiment for unlabeled visual landmarks. We split and merge groups of landmarks to evaluate the impact of group number on the location error of unlabeled landmarks. We also conduct indoor localization of users with labeled visual landmarks as in Sextant [11]. To evaluate the localization accuracy with alteration, we conduct a simulation in areas with altered landmarks and compare localization results of users with and without landmark update.

We evaluate the power consumption as follows. We kill all other foreground applications and execute the client to record a video (around 30 frames per second) for 5 minutes with three trial devices: Huawei Mate 9, Xiaomi MI 6, and Samsung C5 Pro. We record the power drop and estimate the average power consumption of recording the video. Furthermore, we upload the collected video to a server through Wi-Fi to evaluate the power consumption with video transmission.

Unless otherwise stated, the threshold of landmark detection network is 0.7 and the resolution of collected videos is 1920×1080. In the food court and the shopping mall, we set $r(C)$ and $r(K)$ to 0.5 and 0.45 based on our empirical studies. Furthermore, we conduct user localization experiments using visual landmarks labeled by VILL. We use the triangulation algorithm to locate users as studied in Sextant.

7.2 Landmark Detection and Alteration Identification

We illustrate detected objects with a pretrained model in Figure 10. We can see that the pretrained model detects known objects, such as a person, a dining table, and an umbrella. However, it does not detect regions such as store logos. This is because store logos are not in the training dataset of the pretrained model and are different from objects in terms of color and texture. Consequently, the domain shift between the CoCo training dataset and our visual landmarks is significant, rendering it difficult to detect and identify visual landmarks with the pretrained model. We have randomly selected around 130 images from our dataset in the shopping mall with visual landmarks. Experimental results show that the pretrained model does not detect visual landmarks in them. Therefore, we need to fine-tune the network to detect visual landmarks more effectively.

We evaluate the impact of region detection threshold on the identification accuracy in Figure 11. It shows that as the threshold value increases, the identification accuracy of visual landmarks increases. This is because the network discards noisy regions with low confidence values and keeps those with high values. Consequently, we enhance the identification accuracy. However, the number of identified visual landmarks decreases with large threshold value. To achieve a tradeoff

(a) (b) (c)

Fig. 10. Illustration of landmark detection with the pretrained Faster R-CNN.

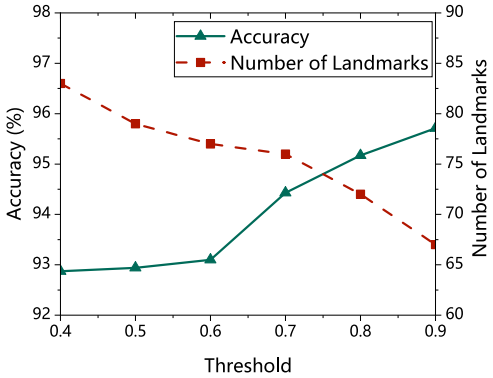


Fig. 11. Landmark identification accuracy vs. threshold (food court).

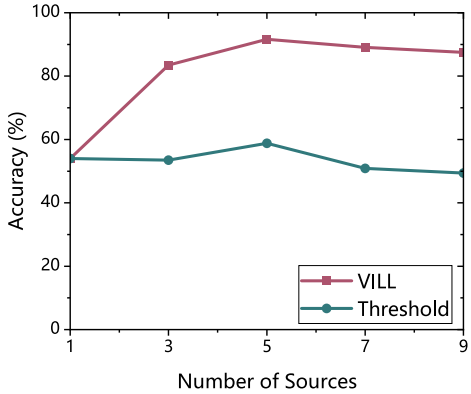


Fig. 12. Alteration identification accuracy vs. number of sources (food court).

between accuracy and the number of detected visual landmarks, we set the threshold value to 0.7 in our experiment.

We present the accuracy of alteration identification with different numbers of sources in Figure 12. It shows that the overall accuracy increases with more sources. This is because we can reduce the impact of opportunistic noises and increase the accuracy with consensus among video sources. As the number of sources is larger than five, the accuracy becomes relatively stable. This is because these sources can provide sufficient information for alteration identification. In the deployment stage, service providers need at least five videos to identify an alteration accurately.

7.3 Visual Landmark Positioning

We present the localization error of visual landmarks in the food court and shopping mall in Figures 13(a) and 13(b), respectively. Experimental results show that our VILL achieves higher accuracy compared with competing schemes. This is because VILL employs an anchor landmark selection strategy to reduce the upper bound of location errors. Furthermore, it incorporates the map constraints to reduce the adverse impact of statistical noises on the location accuracy. The

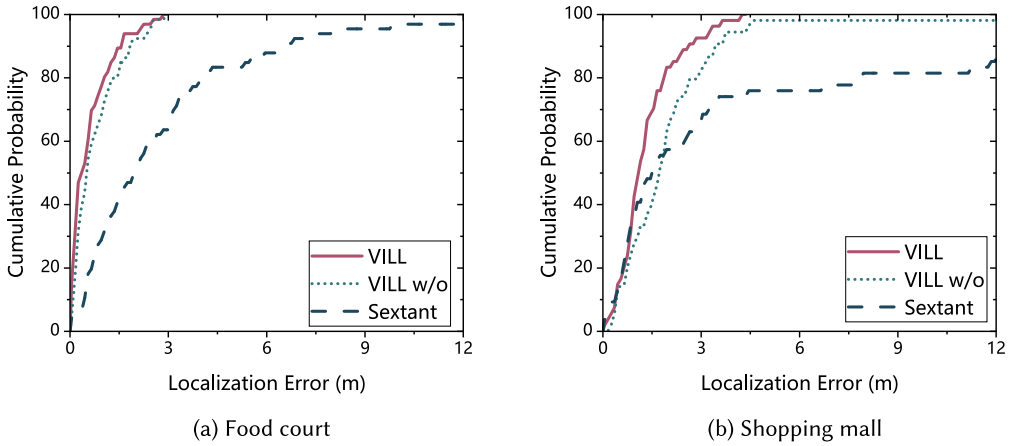


Fig. 13. Localization error of visual landmarks.

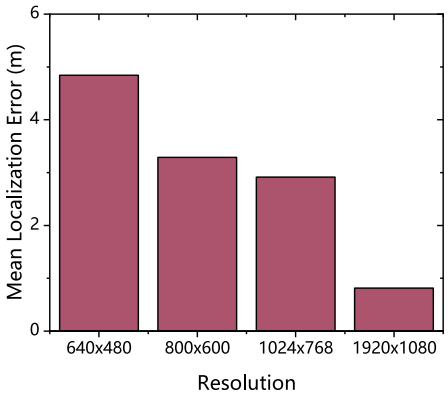


Fig. 14. Localization error vs. resolution (food court).

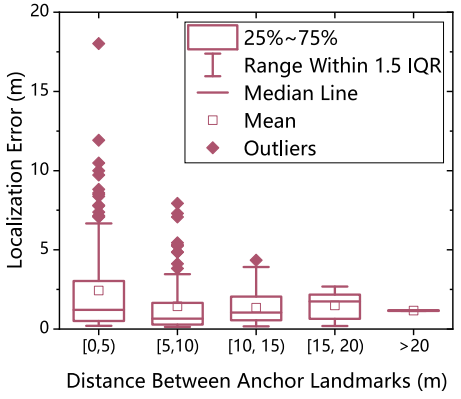


Fig. 15. Localization error vs. anchor distances (food court).

displacement of Sextant, however, is larger compared with VILL due to noisy motion sensors and accumulative errors.

We show the impact of image resolution on the localization error of visual landmarks in Figure 14. It shows that the mean localization error decreases with higher resolution. The reasons are as follows. As the resolution increases, we can extract more distinguishing visual clues from images, thus generating more accurate 3-D models of the environment. Furthermore, landmark detection networks can detect distant visual landmarks accurately, thus increasing accuracy of location estimation for visual landmarks. Consequently, we set the resolution to 1920×1080 in our experiment to achieve sufficient localization accuracy.

We present the localization error of unlabeled visual landmarks with different distances between anchor landmarks in Figure 15. It shows that as the distance increases, the largest and mean localization error of landmarks in the food court decreases. This is because as we increase the distance between anchor landmarks, VILL reduces the adverse impact of statistical noises, thus reducing the localization error of visual landmarks. This demonstrates that the proposed anchor selection strategy can reduce the upper bound of location errors of unlabeled landmarks effectively.

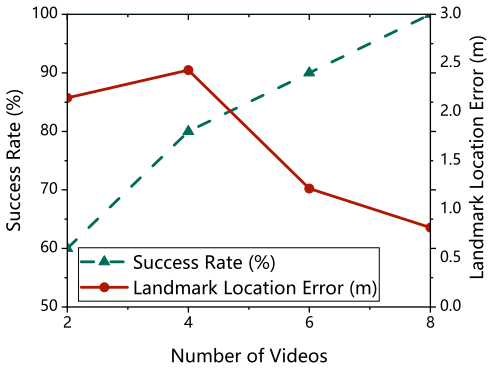


Fig. 16. Success rate and displacement vs. number of videos (food court).

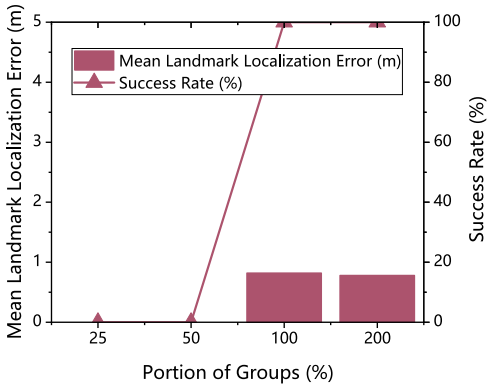


Fig. 17. Mean landmark location error vs. portion of groups.

We show the success rate of landmark labeling and corresponding location error versus the number of videos in Figure 16. It shows that the success rate of landmark labeling increases with the number of videos. This is because we can build more accurate 3-D models of the trial site with more video frames, thus providing accurate spatial correlations and consequently smaller location errors of visual landmarks. In our experiment, we use all videos to label visual landmarks. However, the computational cost in 3-D modeling increases with more videos. To achieve a tradeoff between accuracy and computational cost, one can collect six videos for each group of landmarks.

We present the impact of group number on the location estimation of visual landmarks as well as the success rate in Figure 17. As we reduce the group number by merging remote groups or those separated by textureless areas, the success rate of landmark localization is low. This is because VisualSfM cannot build a point cloud that covers distant groups due to sparse visual features. Consequently, we cannot estimate their locations on the floor plan. As we increase the group number by splitting existing groups into smaller ones, the overall location error of visual landmarks is similar (with marginal decrease). This is because VILL introduces anchor landmark selection, reducing the error bound of unlabeled visual landmarks. Consequently, the accuracy of landmarks does not change significantly.

We illustrate the localization error of users with labeled visual landmarks in Figures 18(a) and 18(b). Experimental results show that the localization error with landmarks labeled by VILL is comparable with that by surveyors. This is because we first use graph-based identification approaches to identify distinguishing visual landmarks from a large number of videos. Based on the anchor selection strategy, our VILL can estimate locations of unlabeled visual landmarks accurately. Consequently, the localization error is comparable with landmarks labeled by surveyors.

We show the localization error of users with and without visual landmark update in Figure 19. Experimental results with altered landmarks show that VILL can achieve comparable localization accuracy with those updated by surveyors manually. The reasons are as follows. VILL can identify altered visual landmarks accurately with several videos uploaded by users. Combined with the accurate alteration identification and localization, we can update landmark labels accurately. Therefore, VILL can achieve sufficient accuracy as visual landmarks labeled manually.

7.4 System Overhead

We evaluate the time consumption of the proposed VILL in each processing step as follows. The time consumption of uploading a video to the server depends on the file size and network

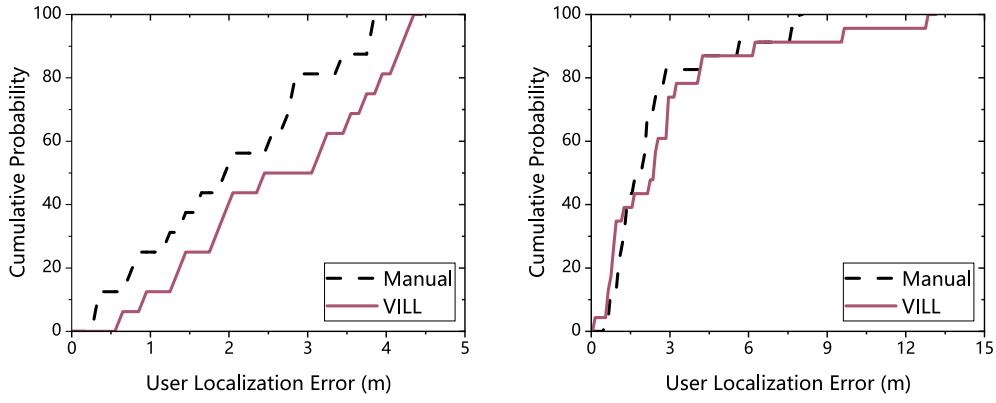


Fig. 18. Localization error of users with visual landmarks.

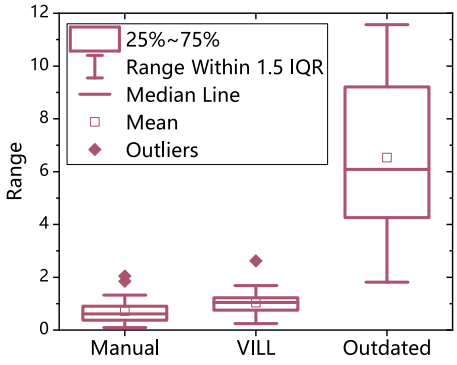


Fig. 19. User localization error with and without update of visual landmarks (food court).

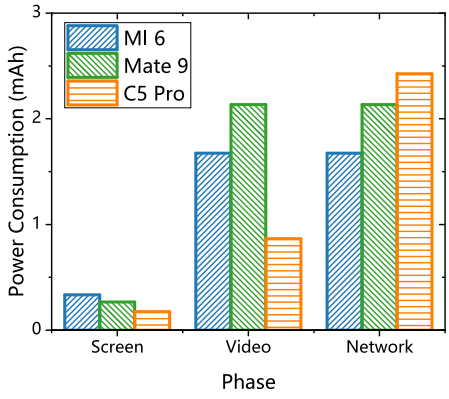


Fig. 20. Power consumption of different trial devices.

bandwidth. For a typical video with 9 seconds (268 frames, 18 MB), the time consumption in video uploading in a 100-Mbps network is around 1.5 seconds. In the landmark identification stage, it takes around 19.8 seconds to process this video in our server. Therefore, it takes around 0.07 seconds (=19.8 ÷ 268) to detect candidate landmark regions in an input frame. As for landmark identification, it takes around 7 seconds using our graph-based identification algorithm. Then, we evaluate the time consumption in the landmark localization stage. The construction of the 3-D model takes around 958 seconds. For a group with nine visual landmarks (two anchor visual landmarks and seven unlabeled ones), it takes 12.2 seconds to locate seven unlabeled visual landmarks. To summarize, the time consumption of landmark labeling with nine visual landmarks is around 997 seconds (=19.8 + 7 + 958 + 12.2). Finally, the time consumption in alteration identification is around 0.011 second. Please note that the landmark labeling is conducted offline, which does not incur any additional overhead for ordinary users.

We evaluate the survey reduction in terms of visual landmarks to be labeled. In the food court, as we have 26 anchor landmarks and 66 unlabeled visual landmarks, the time consumption saved by our VILL is 71.7% (=66 ÷ (26 + 66)). In the shopping mall, we have 28 anchor landmarks and

54 unlabeled landmarks. Consequently, the time consumption saved by VILL is 65.9% (=54 ÷ (28 + 54)). Consequently, we reduce the time consumption by at least 65.9% in the construction stage. Furthermore, as our VILL leverages user-collected videos and identifies altered visual landmarks in the server, surveyors do not have to survey the trial site regularly to identify altered visual landmarks. Therefore, we reduce the time consumption in the update stage significantly.

We illustrate the power consumption with different trial devices in Figure 20. The power consumption with the screen on for around 6 seconds is less than 0.5 mAh. Meanwhile, it takes around 1 to 2 mAh to collect a video clip for 6 seconds. The power consumption of uploading the collected video to a server through Wi-Fi is around 2 mAh. Therefore, the overall power consumption of taking the video and uploading it to a remote server is less than 5 mAh, which is marginal compared with the battery capacity of the state-of-the-art smartphones (around 4,000 mAh).

8 DISCUSSION

In this article, we mainly focus on the labeling of visual landmarks in an indoor environment. We have conducted extensive experiments in two different trial sites. Experimental results demonstrate that the proposed VILL achieves effectiveness and reduces on-site survey effort by surveyors. Although we have addressed the essential problem of landmark labeling and have achieved accuracy and effectiveness, a few more problems remain to be addressed (which are not the focus of this artile).

Transmission Overhead of Videos. VILL is based on the client-server system architecture, where the client program records videos and sends them to a remote server for visual landmark identification. This may incur delay for the client program due to long videos or network congestion. Motivated by the recent advances of mobile devices and algorithms, it is possible to study key frame selection in mobile devices. Afterward, the client program uploads key frames instead of videos to the server, reducing the transmission delay.

Location Privacy of Volunteers. The incorporation of mobile crowdsensing in recording and uploading video clips enables the fast construction and update of a visual landmark database. However, using user-collected videos, backend services can infer the user location or trajectory, incurring severe privacy concerns of volunteers. Although much research effort has been devoted to the protection of location privacy, recent works show that they may be compromised by jointly considering visual representation, correlation, and annotation [51]. In the future, we plan to study more robust techniques to protect the location privacy of volunteers.

Visual Landmark Selection in an Outdoor Environment. In this article, we select store logos as visual landmarks, mainly because they are visually distinguishing from the background and relatively stable (in terms of weeks). In addition to store logos, we also select advertisements on the wall and sculptures as visual landmarks in the food court because they are also visually distinguishing from the background. Therefore, we can also select posters and advertisements as landmarks outdoors. In the future, we plan to conduct more experiments to evaluate the performance with different types of visual landmarks in outdoor scenarios.

9 CONCLUSION

We proposed VILL, an efficient visual landmark labeling approach by camera sweeping. To identify visual landmarks from unstructured videos, we proposed a graph-based visual similarity representation structure, where we identified various visual landmarks accurately and adaptively without knowing their numbers by finding subgraphs. To reduce the location error of unlabeled landmarks, we proposed an anchor selection strategy, where we selected anchor landmarks to guide the location estimation of unlabeled ones. Finally, we proposed an alteration identification model, where we identified altered visual landmarks accurately by considering the user consensus. We

implemented VILL and conducted extensive experiments in a crowded food court and a spacious shopping mall. Experimental results showed that our VILL can reduce the survey effort by at least 65.9% with comparable localization accuracy of visual landmarks.

REFERENCES

[1] Mohamed Abdelaal, Daniel Reichelt, Frank Dürr, Kurt Rothermel, Lavinia Runceanu, Susanne Becker, and Dieter Fritsch. 2018. ComNSense: Grammar-driven crowd-sourcing of point clouds for automatic indoor mapping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (March 2018), Article 1, 26 pages.

[2] Yasin Almalioglu, Muhamad Risqi U. Saputra, Pedro P. B. de Gusmão, Andrew Markham, and Niki Trigoni. 2019. GANVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In *Proceedings of the 2019 International Conference on Robotics and Automation*. IEEE, Los Alamitos, CA, 5474–5480.

[3] Yasin Almalioglu, Mehmet Turan, Muhamad Risqi U. Saputra, Pedro P. B. de Gusmão, Andrew Markham, and Niki Trigoni. 2022. SelfVIO: Self-supervised deep monocular visual-inertial odometry and depth estimation. *Neural Networks* 150 (2022), 119–136.

[4] Heba Aly, Anas Basalamah, and Moustafa Youssef. 2017. Automatic rich map semantics identification through smartphone-based crowd-sensing. *IEEE Transactions on Mobile Computing* 16, 10 (2017), 2712–2725.

[5] Roshan Ayyalasomayajula, Aditya Arun, Chenfeng Wu, Sanatan Sharma, Abhishek Rajkumar Sethi, Deepak Vasisht, and Dinesh Bharadia. 2020. Deep learning based wireless localization for indoor navigation. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. ACM, New York, NY, Article 17, 14 pages.

[6] Yuan Chen, Keiko Katsuragawa, and Edward Lank. 2020. Understanding viewport- and world-based pointing with everyday smart devices in immersive augmented reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–13.

[7] Junyoung Choi, Gyujin Lee, Sunghyun Choi, and Saewoong Bahk. 2022. Smartphone based indoor path estimation and localization without human intervention. *IEEE Transactions on Mobile Computing* 21, 2 (2022), 681–695.

[8] Erqun Dong, Jingao Xu, Chenshu Wu, Yunhao Liu, and Zheng Yang. 2019. Pair-Navi: Peer-to-peer indoor navigation with mobile visual SLAM. In *Proceedings of the IEEE Conference on Computer Communications*. IEEE, Los Alamitos, CA, 1189–1197.

[9] Jiang Dong, Marius Noreikis, Yu Xiao, and Antti Ylä-Jääski. 2019. ViNav: A vision-based indoor navigation system for smartphones. *IEEE Transactions on Mobile Computing* 18, 6 (2019), 1461–1475.

[10] Liang Dong, Jingao Xu, Guoxuan Chi, Danyang Li, Xinglin Zhang, Jianbo Li, Qiang Ma, and Zheng Yang. 2021. Enabling surveillance cameras to navigate. *ACM Transactions on Sensor Networks* 17, 4 (Sept. 2021), Article 35, 20 pages.

[11] Ruipeng Gao, Yang Tian, Fan Ye, Guojie Luo, Kaigui Bian, Yizhou Wang, Tao Wang, and Xiaoming Li. 2016. Sextant: Towards ubiquitous indoor localization service by photo-taking of the environment. *IEEE Transactions on Mobile Computing* 15, 2 (Feb. 2016), 460–474.

[12] Ruipeng Gao, Xuan Xiao, Weiwei Xing, Chi Li, and Lei Liu. 2022. Unsupervised learning of monocular depth and ego-motion in outdoor/indoor environments. *IEEE Internet of Things Journal* 9, 17 (2022), 16247–16258.

[13] Ruipeng Gao, Mingmin Zhao, Tao Ye, Fan Ye, Guojie Luo, Yizhou Wang, Kaigui Bian, Tao Wang, and Xiaoming Li. 2016. Multi-story indoor floor plan reconstruction via mobile crowdsensing. *IEEE Transactions on Mobile Computing* 15, 6 (June 2016), 1427–1442.

[14] Ruipeng Gao, Bing Zhou, Fan Ye, and Yizhou Wang. 2019. Fast and resilient indoor floor plan construction with a single user. *IEEE Transactions on Mobile Computing* 18, 5 (May 2019), 1083–1097.

[15] Yongbin Gao, Fangzheng Tian, Jun Li, Zhijun Fang, Saba Al-Rubaye, Wei Song, and Yier Yan. 2022. Joint optimization of depth and ego-motion for intelligent autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*. Early access, March 24, 2022. https://doi.org/10.1109/TITS.2022.3159275

[16] Fuqiang Gu, Xuke Hu, Milad Ramezani, Debaditya Acharya, Kourosh Khoshelham, Shahrokh Valaee, and Jianga Shang. 2019. Indoor localization improved by spatial context—A survey. *ACM Computing Surveys* 52, 3 (July 2019), Article 64, 35 pages.

[17] Maya Gupta, Ali Abdolrahmani, Emory Edwards, Mayra Cortez, Andrew Tumang, Yasmin Majali, Marc Lazaga, et al. 2020. Towards more universal wayfinding technologies: Navigation preferences across disabilities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–13.

[18] Farzam Hejazi, Katarina Vuckovic, and Nazanin Rahnavard. 2021. DyLoc: Dynamic localization for massive MIMO using predictive recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Communications*. IEEE, Los Alamitos, CA, 1–9.

[19] Chao Huang, Haoran Yu, Jianwei Huang, and Randall A. Berry. 2021. Strategic information revelation in crowdsourcing systems without verification. In *Proceedings of the IEEE Conference on Computer Communications*. IEEE, Los Alamitos, CA, 1–10.

[20] Gang Huang, Zhaozheng Hu, Jie Wu, Hanbiao Xiao, and Fan Zhang. 2020. WiFi and vision-integrated fingerprint for smartphone-based self-localization in public indoor scenes. *IEEE Internet of Things Journal* 7, 8 (2020), 6748–6761.

[21] Shaocheng Jia, Xin Pei, Xiao Jing, and Danya Yao. 2022. Self-supervised 3D reconstruction and ego-motion estimation via on-board monocular video. *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (2022), 7557–7569.

[22] Hongbo Jiang, Wenping Liu, Guoyin Jiang, Yufu Jia, Xingjun Liu, Zhicheng Lui, Xiaofei Liao, Jing Xing, and Daibo Liu. 2021. Fly-Navi: A novel indoor navigation system with on-the-fly map generation. *IEEE Transactions on Mobile Computing* 20, 9 (2021), 2820–2834.

[23] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. 2019. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, Los Alamitos, CA, 8420–8429.

[24] Danyang Li, Jingao Xu, Zheng Yang, Yumeng Lu, Qian Zhang, and Xinglin Zhang. 2021. Train once, locate anytime for anyone: Adversarial learning based wireless localization. In *Proceedings of the IEEE Conference on Computer Communications*. IEEE, Los Alamitos, CA, 1–9.

[25] Danyang Li, Jingao Xu, Zheng Yang, Chenshu Wu, Jianbo Li, and Nicholas D. Lane. 2022. Wireless localization with spatial-temporal robust fingerprints. *ACM Transactions on Sensor Networks* 18, 1 (Oct. 2022), Article 15, 23 pages.

[26] Qing Li, Jiasong Zhu, Tao Liu, Jon Garibaldi, Qingquan Li, and Guoping Qiu. 2017. Visual landmark sequence-based indoor localization. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery (GeoAI'17)*. ACM, New York, NY, 14–23.

[27] Tao Li, Dianqi Han, Yimin Chen, Rui Zhang, Yanchao Zhang, and Terri Hedgpeth. 2020. IndoorWaze: A crowdsourcing-based context-aware indoor navigation system. *IEEE Transactions on Wireless Communications* 19, 8 (2020), 5461–5472.

[28] Manni Liu, Jialuo Du, Qing Zhou, Zhichao Cao, and Yunhao Liu. 2021. EyeLoc: Smartphone vision-enabled plug-n-play indoor localization in large shopping malls. *IEEE Internet of Things Journal* 8, 7 (2021), 5585–5598.

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, Los Alamitos, CA, 10012–10022.

[30] David G. Lowe. 2004. Image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 2 (Nov. 2004), 91–110.

[31] Qun Niu, Mingkuan Li, Suining He, Chengying Gao, S. H. Gary Chan, and Xiaonan Luo. 2019. Resource-efficient and automated image-based indoor localization. *ACM Transactions on Sensor Networks* 15, 2 (Feb. 2019), Article 19, 31 pages.

[32] Meng-Shiuan Pan and Kuan-Ying Li. 2021. ezNavi: An easy-to-operate indoor navigation system based on pedestrian dead reckoning and crowdsourced user trajectories. *IEEE Transactions on Mobile Computing* 20, 2 (2021), 488–501.

[33] Shaifali Parashar, Mathieu Salzmann, and Pascal Fua. 2020. Local non-rigid structure-from-motion from diffeomorphic mappings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 2059–2067.

[34] Milan D. Redžić, Christos Laoudias, and Ioannis Kyriakides. 2020. Image and WLAN bimodal integration for indoor user localization. *IEEE Transactions on Mobile Computing* 19, 5 (2020), 1109–1122.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15)*. 91–99.

[36] Kari Sentz and Scott Ferson 2002. *Combination of Evidence in Dempster-Shafer Theory*. Vol. 4015. Sandia National Laboratories, Albuquerque, NM.

[37] Xingfa Shen, Chuang Li, Weijie Chen, Yongcai Wang, and Quanbo Ge. 2022. Transition model-driven unsupervised localization framework based on crowd-sensed trajectory data. *ACM Transactions on Sensor Networks* 18, 2 (Jan. 2022), Article 26, 21 pages.

[38] Xiaoqiang Teng, Deke Guo, Yulan Guo, Xiang Zhao, and Zhong Liu. 2018. SISE: Self-updating of indoor semantic floorplans for general entities. *IEEE Transactions on Mobile Computing* 17, 11 (2018), 2646–2659.

[39] Liang Wang, Dingqi Yang, Zhiwen Yu, Qi Han, En Wang, Kuang Zhou, and Bin Guo. 2023. Acceptance-aware mobile crowdsourcing worker recruitment in social networks. *IEEE Transactions on Mobile Computing* 22, 2 (2023), 634–646.

[40] Tao Wang, Xiaopeng Zhang, Li Yuan, and Jiashi Feng. 2019. Few-shot adaptive faster R-CNN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 7173–7182.

[41] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. 2020. DeepSFM: Structure from motion via deep bundle adjustment. In *Computer Vision—ECCV 2020*. Lecture Notes in Computer Science, Vol. 12346. Springer, 230–247.

[42] Hang Wu, Jiajie Tan, and S.-H. Gary Chan. 2022. Pedometer-free geomagnetic fingerprinting with casual walking speed. *ACM Transactions on Sensor Networks* 18, 1 (Oct. 2022), Article 8, 21 pages.

[43] Han Xu, Zheng Yang, Zimu Zhou, Longfei Shangguan, Ke Yi, and Yunhao Liu. 2016. Indoor localization via multi-modal sensing on smartphones. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, New York, NY, 208–219.

[44] Jingao Xu, Erqun Dong, Qiang Ma, Chenshu Wu, and Zheng Yang. 2021. Smartphone-based indoor visual navigation with leader-follower mode. *ACM Transactions on Sensor Networks* 17, 2 (May 2021), 22 pages.

[45] Yuri D. V. Yasuda, Luiz Eduardo G. Martins, and Fabio A. M. Cappabianco. 2020. Autonomous visual navigation for mobile robots: A systematic literature review. *ACM Computing Surveys* 53, 1 (Feb. 2020), Article 13, 34 pages.

[46] Lotfi A. Zadeh. 1986. A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. *AI Magazine* 7, 2 (1986), 85.

[47] Dian Zhang, Wen Xie, Zexiong Liao, Wenzhan Zhu, Landu Jiang, and Yongpan Zou. 2022. Beyond RSS: A PRR and SNR aided localization system for transceiver-free target in sparse wireless networks. *IEEE Transactions on Mobile Computing* 21, 11 (2022), 3866–3879.

[48] Yifan Zhang and Xinglin Zhang. 2021. Price learning-based incentive mechanism for mobile crowd sensing. *ACM Transactions on Sensor Networks* 17, 2 (June 2021), Article 17, 24 pages.

[49] Yanchao Zhao, Jing Xu, Jie Wu, Jie Hao, and Hongyan Qian. 2020. Enhancing camera-based multimodal indoor localization with device-free movement measurement using WiFi. *IEEE Internet of Things Journal* 7, 2 (2020), 1024–1038.

[50] Siwang Zhou, Yi Lian, Daibo Liu, Hongbo Jiang, Yonghe Liu, and Keqin Li. 2022. Compressive sensing based distributed data storage for mobile crowdsensing. *ACM Transactions on Sensor Networks* 18, 2 (Feb. 2022), Article 25, 21 pages.

[51] Tongqing Zhou, Zhiping Cai, and Fang Liu. 2021. The crowd wisdom for location privacy of crowdsensing photos: Spear or shield? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (Sept. 2021), Article 142, 23 pages.

Received 11 March 2022; revised 30 October 2022; accepted 27 December 2022