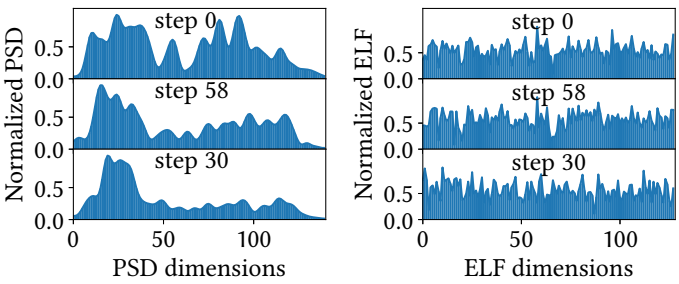
Supplementary File

Wenjie Luo, Qun Song, Zhenyu Yan, Rui Tan, Guosheng Lin

This document includes the supplemental materials for "Indoor Smartphone SLAM with Acoustic Echoes".

Appendix A
Feature Visualization

A.1 Feature similarity visualization comparison

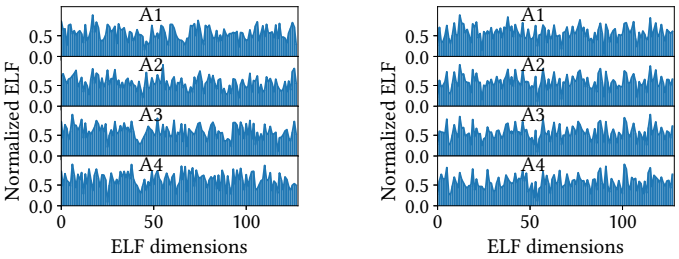


(a) PSDs (b) ELFs

Fig. 1: (a) sim(0,30) = 0.42, sim(0,58) = 0.4 (b) sim(0,30) = 0.22, sim(0,58) = 0.61.

Fig. 1 shows the cosine similarities (denoted by $\mathrm{sim}(\cdot,\cdot)$) using PSDs and ELFs for loop closure steps 0 and 58 and non-loop closure steps 0 and 30. For PSD, $\mathrm{sim}(0,58)$ is 0.4 and $\mathrm{sim}(0,30)$ is 0.42, there are no significant differences. PSD is ineffective to differentiate the loop/non-loop closure data. For ELF, $\mathrm{sim}(0,58)$ is 0.61 and $\mathrm{sim}(0,30)$ is 0.22, ELF can effectively differentiate loop/non-loop closures based on the feature similarity measurement.

A.2 ELF visualization after map superimposition



(a) (b)

Fig. 2: Spot A's ELFs. The average similarity is (a) 0.34 and (b) 0.76 before/after map superimposition.

Fig. 2 visualizes the spot A's ELFs from directions 1 to 4 before/after map superimposition. We calculate the cosine similarites among ELFs from different directions. The ELFs' average similarity is 0.34 before the map superimposition. The similarity increases up to 0.76 after applying the floor-level CL for map superimposition. This result shows that map superimposition is effective in reconciling the ELFs' differences due to phone orientations.

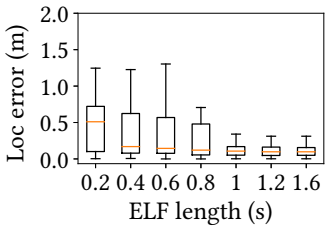


Fig. 3: ELF seq. length.

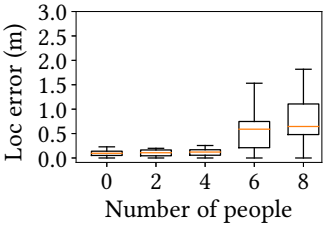


Fig. 4: No. of people.

Appendix B
Sensitivity Analysis for Localization

We conduct experiments mainly in the living room to evaluate the sensitivity of ELF-SLAM to various factors. By default, we consider one-shot localization.

B.1 ELF sequence length.

Fig. 3 shows the localization errors when we vary the length of the ELF sequence used for computing ESS from $0.2\,\mathrm{s}$ to $1.6\,\mathrm{s}$. A boxplot shows the localization error distribution. The horizontal line in each boxplot shows the median. We can see that the localization error decreases with the ELF sequence length and becomes flat when the sequence length is more than $1\,\mathrm{s}$. Note that at human's average walking speed, the duration between two consecutive footsteps is about $0.6\,\mathrm{s}$, which results in an ELF sequence length of $0.6\,\mathrm{s}$ as well. From Fig. 3, at this length setting, the one-shot localization median error is around $0.1\,\mathrm{m}$. Thus, ELF-SLAM performs well when the user walks at a normal speed.

B.2 Nearby moving people.

Human bodies can reflect the excitation chirp and generate irrelevant echoes. We evaluate the impact of the nearby moving people on one-shot localization. Multiple volunteers walk freely in the living room and talk to each other during the localization phase. Fig. 4 shows the localization error versus the number of nearby moving people. The error remains low when the number of people is up to 4. Note that the tested area is about $60\,\mathrm{m}^2$. When there are 6 and 8

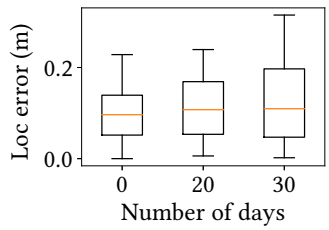


Fig. 5: Aging.

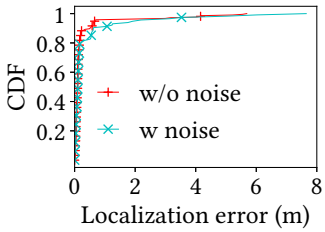
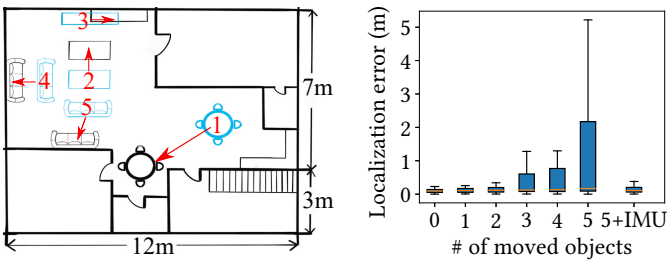


Fig. 6: Audible noises.



(a) (b)

Fig. 7: (a) Movements of furniture objects in a living room; (b) the corresponding localization performance.

moving people, whose crowd density is similar to that in the shopping mall during peak hours, the median localization errors increase to 0.57 m and 0.6 m. Nevertheless, the errors remain at the sub-meter level. Thus, ELF-based localization can tolerate nearby moving people to a certain extent.

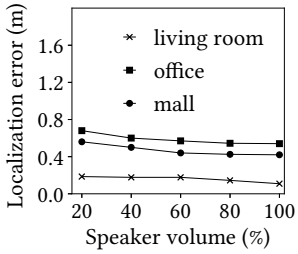


Fig. 8: Speaker volume.

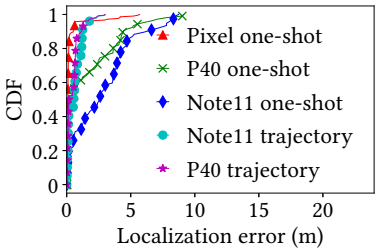


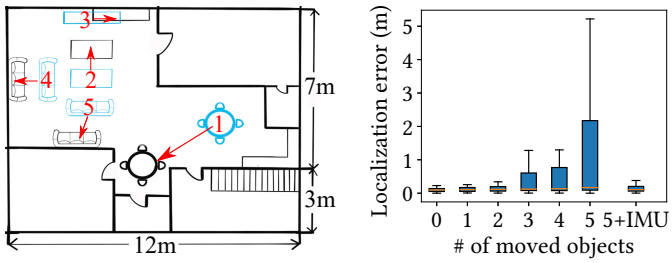
Fig. 9: Hardware heterogeneity.

B.3 Map aging.

We investigate whether the map constructed by ELF-SLAM ages. Specifically, at day 0, we use ELF-SLAM to construct a trajectory map. Then, we evaluate the ELF-based localization performance multiple times during a one-month period. Fig. 5 shows the results. The median localization errors are 0.10 m, 0.11 m, and 0.12 m at day 0, 20, and 30, respectively. This suggests that the constructed map does not have salient aging issue. In practice, a map can be continuously updated using the latest data contributed by users, to mitigate any potential aging issue.

B.4 Audible noises.

We evaluate the robustness of ELF-based localization against audible noises. We use a laptop computer to play video clips of different contents (music, speech, etc) from Youtube to generate the noises. From Fig. 6, the noises have little impact on the localization performance. This is because our system operates within the near-inaudible frequency band. Thus, audible noises have negligible impact on ELF-based localization.



(a) (b)

Fig. 10: (a) Movements of furniture objects in a living room; (b) the corresponding localization performance.

B.5 Space layout changes.

The layout changes of the target space may have impact on the chirp reverberation processes. Thus, we deliberately change the furniture locations in the living room to evaluate such impact. Fig. 10a illustrates how the furniture objects are moved. Specifically, we move five objects including a dining table, a tea table, a TV cabinet, and two sofas. We move one object at a time. Fig. 10b shows the localization error versus the number of moved objects. The error remains low when the number of moved objects is less than 5. When all the 5 objects are moved, the mean localization error increases to 1.3 m. If we apply the trajectory localization, the mean localization error decreases to 0.3 m as labeled by "5+IMU" in Fig. 10b. Therefore, the trajectory localization improves the robustness of ELF-based localization against the layout changes. In practice, a map can be continuously updated using the latest data to mitigate such impact..

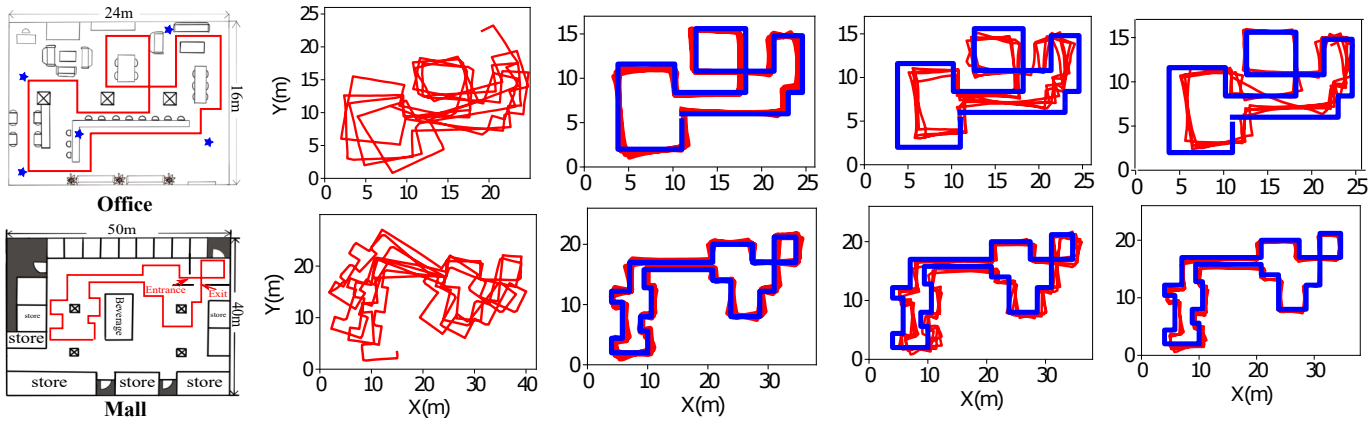
B.6 Speaker volume.

As pets and human infants may have wider hearing limits [1], they may perceive the chirps emitted from the smartphone. To avoid annoyance to them, we evaluate the localization with various settings for the smartphone's loudspeaker volume in emitting the chirps. Fig. 8 shows the results. When the volume decreases from 100% (i.e., the highest volume) to 20%, the median localization errors in the living room, office, and shopping mall increase from 0.1 m, 0.54 m, and 0.42 m to 0.18 m, 0.68 m, and 0.56 m, respectively. Note that with 20% loudspeaker volume, on the audible frequency band, the smartphone's sound is soft and becomes nearly imperceptible in environments with normal noise levels. Thus, ELF-SLAM maintains sub-meter accuracy when the chirp emission is soft.

B.7 Smartphone hardware heterogeneity.

The microphone hardware heterogeneity can cause domain shifts for speech recognition [2]. To evaluate the impact of smartphone hardware heterogeneity on ELF-SLAM, we conduct experiments using three smartphones, i.e., Google Pixel 4, Huawei P40 Pro, and Redmi Note11. We use Pixel for map construction and all three smartphones for localization performance evaluation. Fig. 9 shows the results.

The median localization error in localizing Pixel is only 0.1 m. The median errors in localizing P40 and Note11 increase to 0.5 m and 2.54 m, respectively. The hardware heterogeneity can be a primary reason for the performance



(a) Floor plans (b) IMU (c) ELF (d) Wi-Fi (e) Geomagnetism

Fig. 11: Floor plans and trajectory reconstruction results in shopping mall and office.

drops, echoing the study [2]. Note that the Pixel and P40 have similar list prices, while the Note11 is about $3.5\times$ cheaper. This price comparison is consistent with the observation that Note11 experiences more performance drop than P40. We also evaluate the trajectory localization on the three smartphones. From the results labeled with "trajectory" in Fig. 9, for P40 and Note11, the median localization errors decrease to $0.2\,\mathrm{m}$ and $0.5\,\mathrm{m}$, respectively. Thus, the trajectory localization largely mitigates the negative impact of audio hardware heterogeneity. This result has the following two implications. First, inertial sensing, although suffering long-run drifts, provides important information for localization. Thus, fusing the results of inertial sensing and echo sensing increases the system's robustness. Second, because IMUs are in general low-cost, the three phones' IMUs may be of similar qualities.

Appendix C
Mapping Performance

Mapping performance of three modalities in the shopping mall and the office is shown in Fig. 11

References

[1] "Hearing range," 2022, https://en.wikipedia.org/wiki/Hearing_range.

[2] A. Mathur, T. Zhang, S. Bhattacharya, P. Velickovic, L. Joffe, N. D. Lane, F. Kawsar, and P. Lió, "Using deep data augmentation training to address software and hardware heterogeneities in wearable and smartphone sensing devices," in *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 2018, pp. 200–211.