Preserving Privacy in Mobile Spatial Computing

Nan Wu
George Mason University
Fairfax, Virginia, USA
nwu5@gmu.edu

Ruizhi Cheng
George Mason University
Fairfax, Virginia, USA
rcheng4@gmu.edu

Songqing Chen
George Mason University
Fairfax, Virginia, USA
sqchen@gmu.edu

Bo Han
George Mason University
Fairfax, Virginia, USA
bohan@gmu.edu

ABSTRACT

Mapping and localization are the key components in mobile spatial computing to facilitate interactions between users and the digital model of the physical world. To enable localization, mobile devices keep capturing images of the real-world surroundings and uploading them to a server with spatial maps for localization. This leads to privacy concerns on the potential leakage of sensitive information in both spatial maps and localization images (*e.g.,* when used in confidential industrial settings or our homes). Motivated by the above issues, we present a holistic research agenda in this paper for designing principled approaches to preserve privacy in spatial mapping and localization. We introduce our ongoing research, including learning-assisted noise generation to shield spatial maps, distributed architecture with intelligent aggregation to protect localization images, and end-to-end privacy preservation with fully homomorphic encryption. We also discuss the technical challenges, our preliminary results, and open research problems in those areas.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Mixed / augmented reality**; • **Human-centered computing** → **Mobile computing**; • **Security and privacy** → **Privacy protections**.

KEYWORDS

Mobile Spatial Computing, 3D Spatial Maps, Image-based Localization, Privacy Leakage, Structure from Motion, Fully Homomorphic Encryption

ACM Reference Format:

Nan Wu, Ruizhi Cheng, Songqing Chen, and Bo Han. 2022. Preserving Privacy in Mobile Spatial Computing. In *32nd edition of the Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV '22), June 17, 2022, Athlone, Ireland*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3534088.3534343

1 INTRODUCTION

Mobile spatial computing is an emerging computing paradigm that takes users' physical actions (*e.g.,* head and body movements, gestures, and speech) as input and their visual cortex as the display to interact with their perceived 3D space. It advances the original concept that was defined as "*human interaction with a machine in which the machine retains and manipulates referents to real objects and spaces*" [13] with augmented and mixed reality (AR/MR) headsets such as Microsoft HoloLens and Magic Leap One, by moving beyond the confines of 2D screens [20, 22] More broadly, it refers to "*computing in spatial, temporal, spatiotemporal spaces across both geographic and non-geographic domains*" [34].

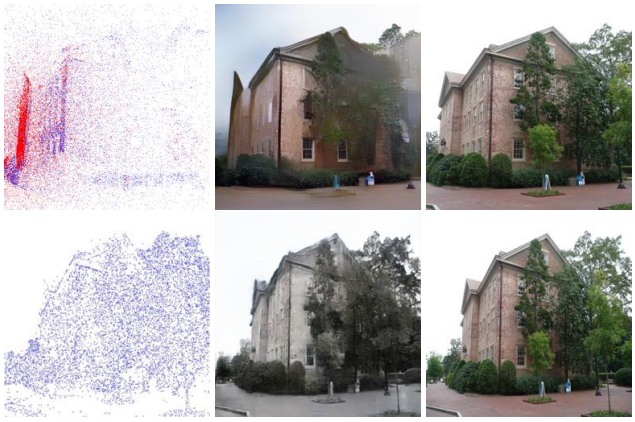

Figure 1: Reconstruction of original scene from a spatial map (top) and localization image from its visual features (bottom). Three columns show the visual features (left), reconstruction results (middle), and the ground truth (right).

To better support AR/MR applications, spatial computing relies on innovations in vision-based spatial mapping and 6DoF (six degrees of freedom) localization. Take Facebook/Meta's Project Aria [8] as an example. It utilizes LiveMaps, a spatial map that consists of sparse 3D points (*i.e.,* a point cloud) with visual feature descriptors, to store explicit geometric information and semantic knowledge of the surrounding environment in volumetric views. Widely-used methods for spatial mapping include structure from motion (SfM) [1, 33] and visual SLAM (simultaneous localization and mapping) [11, 41], which processes images that capture the physical world. During localization, a mobile device uploads its camera-view image to a server that extracts 2D features from the image and matches them with the 3D features in the spatial map to estimate the device's 6DoF pose. This method is referred to as image-based localization [30]. To preserve privacy in localization images, a straightforward method is to let the mobile device extract features and upload them to the server for localization, instead of directly uploading localization images.

However, both the spatial maps and visual features extracted from localization images may still contain sensitive information about users' surrounding environment, leading to enormous privacy concerns. For instance, as shown in Figure 1, once attackers

NOSSDAV '22, June 17, 2022, Athlone, Ireland

ACM ISBN 978-1-4503-9383-6/22/06.
https://doi.org/10.1145/3534088.3534343

or malicious users get access to the visual data, they can reconstruct the original images/scenes with good quality (*i.e.,* a high degree of accuracy) [5, 21, 25, 39]. Thus, this may result in security attacks when AR/MR applications get more popular and are used in life-critical situations (*e.g.,* on a battlefield or in a surgical operating room). Moreover, an adversary can even modify the spatial maps and/or the uploaded features (*e.g.,* man-in-the-middle attacks) and make the localization result significantly deviate from the ground truth. The above issues call for robust mechanisms to shield countless malicious attacks and protect user privacy. Hence, our long-term research goal is to *design principled approaches to simultaneously preserve privacy and defend against security attacks* for spatial mapping and localization in mobile spatial computing.

In this paper, we present a holistic research agenda to make an initial step towards the above ambitious goal by exploring privacy-preserving schemes to protect spatial maps and visual features for localization. We aim to prevent the recovery of scene/image details by answering the following three questions. (1). How to preserve sensitive information in spatial maps? (2). How to protect user privacy in localization images/features? (3). How to provide practical end-to-end protection for spatial mapping and localization with some advanced encryption techniques?

We face several unique challenges when investigating the above three research problems. First, privacy-preserving schemes should not (remarkably) affect localization performance (*e.g.,* localization accuracy and time). Second, when involving mobile devices, those schemes should be lightweight and balance the tradeoff between computation/communication overhead and localization performance. Finally, it is far from trivial to integrate the proposed solutions into existing AR/MR systems and make them practical by scaling to thousands of users.

Towards tackling the above technical challenges, we make the following contributions in this paper.

- We propose to judiciously add perturbation noises to spatial maps for preventing the recreation of original scenes. Our initial exploration reveals the tradeoff between the degree of privacy-preserving for spatial mapping and the increased localization time, which motivates the adoption of adversarial examples [35] that jointly consider those factors in the design of loss function. Although similar techniques have been utilized in robust object classification [3, 24, 40], our problem is more challenging due to the balance of localization time, localization accuracy, and privacy preservation.
- To protect sensitive information in localization images, we propose a distributed architecture that splits the extracted visual features and sends them to multiple randomly selected and independent servers for localization. This approach is based on the observation that reducing the number of available features will drastically degrade the quality of reconstructed images. However, doing this will reduce the localization accuracy. We will study efficient algorithms to minimize the loss of localization accuracy by effectively aggregating results from distributed servers.
- We propose to take advantage of fully homomorphic encryption (FHE) [12] to offer end-to-end protection of sensitive data from privacy leakage and malicious attacks. By enabling operations on encrypted data, FHE can efficaciously preserve user privacy through the encryption of both spatial maps and visual features for localization. We present an initial design and our preliminary results for integrating FHE into existing spatial mapping and localization pipelines, and identify a few key challenges including the reduction of localization error and computation latency caused by FHE.

2 BACKGROUND

Spatial Map Construction. SfM is a widely-used technique for building spatial maps, by utilizing a set of unstructured 2D images to create the 3D model of a scene or an object. The SfM pipeline first extracts features from the images and determines which images are overlapped with each other using those features. An extracted feature has its location in the image (*i.e., x* and *y* pixel coordinates) and a feature descriptor (*e.g.*, a vector with 128 dimensions for SIFT [18]). With corresponding features in pairs of overlapped images, SfM then uses triangulation to calculate the 3D coordinates of those features and the relative pose of those images. Finally, it employs bundle adjustment to refine image poses and 3D point coordinates since pose estimation errors may arise from noisy matches and imprecise calibration. As a result, *a feature in a spatial map consists of its 3D position and a feature descriptor extracted from the images.*

Image-based Localization. There are two key steps in 6DoF image-based localization. It first finds the best matching 3D features from the spatial map of the 2D features extracted from a localization image with the distance of the feature descriptors. It then performs the 6DoF localization of the target image using the P*n*P (Perspective-*n*-Point) method [10] with the set of 2D positions of *n* localization features and the 3D locations of their matched features in the map. Using the camera's intrinsic matrix that is included in the EXIF (exchangeable image file format) metadata of the images, it finally calculates the camera's 6DoF pose (*i.e.,* extrinsic parameters) based on the perspective projection model [36]. P*n*P is usually applied with RANSAC (random sample consensus) [10] to make the localization result robust to outliers in the matched feature pairs.

Sensitive Information Leakage. When mobile applications that involve spatial mapping and localization are used in places such as homes or confidential workspace, they raise significant privacy concerns due to the sensitive visual information that can be revealed from images for map construction, the maps themselves, and visual features in localization images, especially for the mapping process that usually works in the background (*i.e.,* users may not be fully aware of it). It is already known that one can reconstruct the original image fairly well from its visual features, no matter whether they are hand-crafted or generated by deep learning models [5, 21, 39]. Moreover, recent work demonstrated that it is feasible to recover scene details even from the sparse point cloud in a spatial map [25]. Our proposed work is motivated by sensitive information leakage in spatial mapping and localization of mobile spatial computing.

Privacy & Security for Vision-based Applications. Privacy, security, and safety issues have always been a key concern for vision-based applications such as AR/MR [16, 17, 27, 28] and video analytics [2, 26, 38]. For example, Recognizer [16] is an OS abstraction that offers fine-grained access control to AR applications (*e.g.,* exposing a skeleton, instead of raw sensor data). Sabelman and Lam [28] explored potential hazards that could be caused by head-mounted displays such as Google Glass and Microsoft HoloLens. PECAM [38]

is a system that can not only preserve the privacy of video analytics but also offer a transformation that is securely-reversible. The privacy-enhanced videos can be fully reversed in the case of forensics usage, without requiring any auxiliary data. Privid [2] provides duration-based privacy for video analytics by defining a new notion of differential privacy for this task, which protects sensitive information for a particular duration. Instead, we focus on preserving privacy and defending security attacks in vision-based mapping and localization for mobile spatial computing.

3 RESEARCH AGENDA

In this section, we first present our target threat models. We then introduce two complementary methods to prevent the reconstruction of sensitive information in spatial maps and localization images, respectively. Finally, we propose an end-to-end privacy preservation and security protection framework using FHE.

3.1 Threat Models

The spatial mapping and localization service involves two parties, the client and the server, and the communication between them. Thus, any part of this service could be subject to attacks and/or information leakage. In the first (and least threatening) model, we assume that the localization server is trustworthy but curious. At the same time, we assume the communication between the client and the server is reliable and trustworthy. Our first line of proposed research that strategically adds perturbation noises and distributes visual features to multiple randomly selected and independent localization servers targets this model.

We further consider a stronger threat model, where the server could be compromised and actively probe the information supplied by users, and the information communicated between the clients and the server could be subject to eavesdropping or other attacks (*e.g.,* a man-in-the-middle attack). Our second line of proposed research that utilizes the FHE for end-to-end protection targets this stronger model.

3.2 Preserving Privacy for Spatial Maps

Background. Existing work [25] recovers scene details from a spatial map in three steps, visibility estimation of features, coarse scene reconstruction, and refinement of visual quality. More specifically, it renders the spatial map from a specific viewpoint into a 2D image, feeds the rendered image into a series of neural networks, and outputs a color image of the to-be-recovered scene. To properly render the image, it needs to estimate which 3D features in the spatial map are visible from the given viewpoint and should be projected onto the rendered image. This image is then fed into an image synthesis network for coarse scene reconstruction. The visual quality of the recreated scene is further refined by an adversarial framework. The good quality of the reconstructed scene, as shown in Figure 1 (top), can potentially lead to privacy leakage. The capability of scene reconstruction from spatial maps motivates us to study the following research problem of privacy-preserving spatial mapping.

Problem: How can we preserve sensitive information in spatial maps without affecting localization performance?

Figure 2: An example of a recovered scene from the original spatial map (left) and the same recreated scenes from the maps with 10% added noisy features (middle) and 100% added noisy features (right), compared to the size of the original map.

Our Solution. We propose to judiciously introduce perturbation noises to a spatial map for impeding the reconstruction of original scenes. We duplicate part of the initial visual features according to the noise level (*e.g.,* 10%), and then add Gaussian noises to their positions, RGB values, and feature descriptors. Note that we do not change the original visual features in the spatial map. The rationale behind this method is that the quality of recovered scenes heavily depends on the first step that estimates the visibility of features. If the added noises could be retained during this visibility estimation step and contribute to the reconstruction, they will inevitably create artifacts and degrade the visual quality of recovered scenes. It would even be better if they could occlude some important visible features in the spatial map and hide the scene details, further decreasing the visual quality of the reconstructed scene. We emphasize that our proposed solution has a different threat model compared to noise addition in differential privacy [6, 7]. We aim to shield the sensitive information in the spatial maps in case they may be revealed to malicious users, and we do not seek to protect the localization results in this study. On the other hand, differential privacy aims to shield the confidential information of individuals that could be inferred from queries made by malicious users (*i.e.,* differential privacy protects the query result).

While this method sounds straightforward, there are three key challenges to making it effective. First, the localization accuracy should not be affected by the introduced perturbation noises. Second, while adding more noises could further reduce the visual quality of reconstructed scenes (as shown in Figure 2), we should minimize the number of introduced noisy features. The reason is that after being part of the spatial map, they will increase the number of matches that need to be performed against the features in the localization image and thus prolong the localization time. Third, the denoise of the added perturbations should be non-trivial.

Note that different from robust object classification [3, 24, 40], we seek to add a small amount of noises into the spatial map that will affect the reconstruction conducted by deep neural networks [25] as much as possible and thus reduce the quality of reconstructed scenes for protecting the privacy information. On the other hand, existing adversarial attacks mainly focus on creating false classification results of deep neural networks, where adding small adversarial perturbations into the input leads to misclassification. However, the difference between the original input and the adversarial input is undetectable by a human.

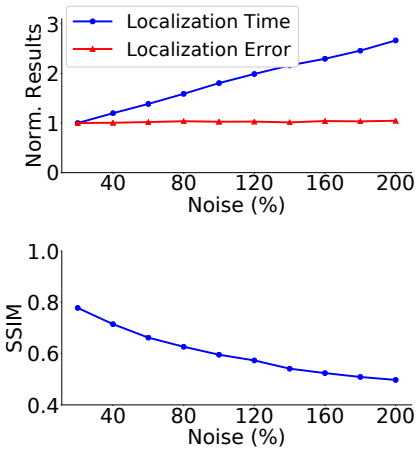

Figure 3: Normalized localization time and errors (top) and visual quality of reconstructed scenes (bottom) caused by different amounts of added noises.

Preliminary Results. To explore the tradeoff between the penalty of injecting noises (*e.g.,* increased localization time and degraded localization accuracy) and the effectiveness of privacy preservation, we conduct a pilot study that adds Gaussian noises to spatial maps. We use the COLMAP [33] SfM pipeline to generate a spatial map with its "South Building" dataset that contains 128 images taken from different angles of a building at UNC-Chapel Hill. We use the pre-trained InvSfM model [25] for scene recreation, and evaluate the performance of our proposed solution on the same dataset.

We measure the impact on localization accuracy and the quality of recreated scenes after adding noisy features into the spatial map as follows. We calculate the position error P_e as the Euclidean distance between an image's estimated position and the ground truth. The orientation error O_e is defined as the minimum rotation angle required to align the estimated rotation and the ground truth [14]. After adding Gaussian noises to the spatial map, we calculate again the position and orientation errors, P_n and O_n. We report the normalized errors, P_n/P_e and O_n/O_e, to demonstrate the impact of adding noises on localization accuracy. We use SSIM [37], the structural similarity metric, to measure the visual quality of recovered scenes from the spatial map with added noises, using the reconstructed ones from the original map as the reference. In this paper, we resize the resolution of localization images to be 1024×768, as the highest resolution that our GPU can support for scene recreation is 1024×1024.

We plot the normalized localization time and errors and the SSIM of recreated scenes in Figure 3. In the two sub-figures, the x-axis shows the percentage of added noises, compared to the size of the original spatial map. As shown in the top sub-figure, while adding noises has a limited impact on localization accuracy, the localization time linearly increases with the number of noisy features as they are used to match against localization features. We plot only the normalized position errors. The orientation errors show a similar trend (*i.e.,* not impacted by the noisy features). The bottom sub-figure indicates that introducing more noises into the spatial map can indeed reduce the visual quality of recovered scenes. The SSIM value drops with more added noisy features. Hence, it is challenging to determine the optimal percentage of added noise by balancing the localization time and the quality of reconstructed scenes.

Discussion. We use the Gaussian noises to explore the research challenges faced by our proposed solution. In practice, we should deliberately add noisy features that cannot be easily denoised. Given the success of adversarial examples [35], we plan to train an adversarial model that can judiciously add perturbation noises to effectively hide sensitive visual information in a spatial map during the reconstruction. A key requirement here is to impede the scene reconstruction by adding as few noisy features as possible, with the goal of limiting its impact on localization time. Moreover, we should consider the potential denoising that could be performed by attackers and add perturbation noises that are difficult to remove. Thus, when building the adversarial model, we should jointly consider the denoising overhead and multiple factors in the loss function, including the localization time and accuracy and the visual quality of recovered scenes.

There are several open challenges to our proposed solution. First, there should be a widely-accepted definition of privacy leakage in spatial maps. Different users may have different perceptions of privacy leakage. The sensitive level of a piece of information may differ for different applications and under different scenarios. Second, we should have a well-defined metric to quantify the effectiveness of privacy-preserving schemes. The SSIM metric is applied to the entire image. In reality, users may care more about the prevention of reconstructing points of interest that can be used to identify a location. Third, while most existing spatial mapping and localization systems leverage SfM frameworks that utilize hand-crafted features (*e.g.,* SIFT) [33], recent proposals such as hloc [29] benefit from features extracted by deep-learning models. We will explore the feasibility of applying our proposed scheme to learning-based spatial map construction.

3.3 Privacy-preserving Localization

Background. Besides preserving privacy in spatial maps, we also need to prevent the leakage of sensitive information in localization images. The state-of-the-art [5, 21, 39] inverts 2D visual features extracted from an image into the original image through various methods such as convolution networks and deep generative models. Although adding noises to the features may decrease the visual quality of reconstructed images, applying this method to localization features is more challenging than doing this to the features in a spatial map. The reason is that adding these noises will not only unavoidably reduce localization accuracy but also increase communication overhead. Furthermore, the noises should be added on mobile devices before sending the features for real-time localization, making it infeasible to utilize the above method that benefits from adversarial models due to its high computation overhead. Thus, we will investigate the following research problem for privacy-preserving localization.

Problem: How can we preserve sensitive information in localization images with mobile-friendly methods?

Our Solution. We explore a systems approach for preserving sensitive visual data by splitting the extracted features into N groups,

Figure 4: An example of a reconstructed image from its features (left) and the same reconstructed image from 50% of its features (middle) and 10% of its features (right).

distributing them to N out of M ($N \ll M$) randomly selected and independent servers for simultaneous localization, and intelligently aggregating the outputs to get the final localization result. For example, if we randomly and uniformly divide the features into 5 groups of roughly the same size, each group will contain about 20% of the features. In this case, the localization servers have only partial information of the original image (*e.g.,* 20% feature), making the reconstruction difficult or even infeasible. We emphasize that we assume these servers are independent and thus do not have the incentive to collude and merge the received visual features for reconstructing the localization image.

Preliminary Results. This approach is motivated by the observation that the visual quality of the reconstructed image will dramatically degrade when reducing the number of available features. We train the SIFT-Reconstruction model [39] with 64 images of the "South Building" dataset (Section 3.2) and 3,583 images of the large Aachen dataset [31, 32] that contains 4,479 images to address the potential overfitting issue. We show an example of a reconstructed image with this model in Figure 4 (left). The middle and right sub-figures in Figure 4 are the reconstructed images using 50% and 10% features, respectively. As we can see from this figure, if a server has only 10% features extracted from the localization image, the reconstructed image is blurred with a low visual quality. When using the ground truth image as the reference to measure the visual quality, the calculated SSIM value is only 0.28 for the right sub-figure that is recreated with 10% features.

While the distributed scheme above can protect sensitive information in localization images, using few features reduces localization accuracy. In Figure 5, we plot the normalized localization errors of distributing the extracted features to 10 and 20 servers, respectively, compared to the approach that uses all features. For 50% of cases, the normalized position error and the normalized orientation error are higher than 2.2 when using 10 servers. The results are even worse with 20 servers. Thus, the key challenge for this research problem is to design an intelligent aggregation algorithm that can improve localization accuracy. We evaluate the performance of a simple scheme that averages the localization results from individual servers. Although this scheme can improve the localization accuracy for most cases, there are still 10% of localization with normalized errors higher than 1.3. We will study more effective aggregation algorithms that can further boost localization accuracy of our proposed scheme.

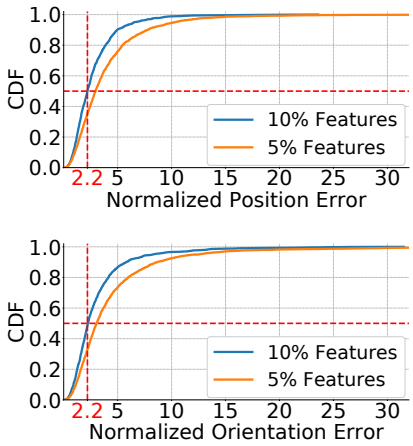

Figure 5: CDF of normalized localization errors using 5% and 10% features, compared to using all available features.

Discussion. Similar to the problem of preserving privacy in spatial maps, there are a few open challenges to protect sensitive information in localization images, such as the definition of privacy leakage, the metrics to quantitatively measure the effectiveness of the proposed privacy-preservation schemes, and the extension to learning-based features. In addition, we will explore how to determine the proper number of groups N that the extracted visual features should be divided into. If N is too small, the reconstructed localization image may still bear a high degree of accuracy (*e.g.,* middle sub-figure in Figure 4 for $N = 2$). On the other hand, if N is too large, the localization accuracy on individual servers will drastically degrade, as shown in Figure 5, making the aggregation difficult. Another open problem is how to effectively break the localization features into groups. We will examine the importance of different features in terms of their contribution to localization accuracy and the recreation of the original image. If there are features that provide more value to localization but less value to reconstruction, we can repeat them in different groups.

3.4 End-to-End Protection with FHE

Background. We explore fully homomorphic encryption (FHE) [12] to shield spatial mapping and localization from privacy leakage and security attacks. By enabling additions and multiplications on encrypted data, FHE is able to defend potential attacks while preserving privacy in spatial maps and localization features. However, there are a number of challenges to adopting FHE for this purpose. First, it is non-trivial to integrate FHE into existing spatial mapping and localization systems, given their complex operations in the pipeline and different types of data to deal with. For example, it may not be practical to encrypt all involved data sources. Second, the current generation of FHE algorithms still cannot handle some of the non-polynomial operations that are required for spatial mapping and localization. Such operations should be replaced with approximation. Moreover, some of the schemes in FHE create only approximate results [4]. This inevitably leads to a decrease in

localization accuracy. Third, FHE is well-known for its poor runtime performance and the inflated size of encrypted data, despite continuous improvements. Enhancing its efficiency and making it feasible to execute on mobile devices and deliver encrypted data over mobile networks is still an unsolved problem. Thus, we will examine the following research problem for utilizing FHE to protect mobile spatial computing.

Problem: How can we provide practical end-to-end protection for spatial mapping and localization using FHE?

Our Solution. We present our initial solutions to some of the aforementioned challenges. We first investigate what data should be encrypted with FHE. Ideally, we should encrypt both the 2D/3D positions and the feature descriptors extracted from localization images and those in spatial maps. However, doing this brings about two issues. First, when finding the matching features in the spatial map for localization features, we should compare the distances of different pairs of feature descriptors and select the best match. While recent proposals such as Pegasus [19] can support the comparison operation, its computation overhead is prohibitively high. Second, the size of FHE-encrypted data is much larger than the original data, as we will show next. By considering the fact that the size of a feature descriptor is much larger than that of the 2D/3D coordinates, we encrypt only the 2D/3D positions of the features and leave their descriptors as plain-text to make feature-matching feasible and practical. This will significantly reduce the visual quality of recreated images from the extracted features [39].

We next study the technical challenges of estimating the 6DoF pose with FHE-encrypted data. As mentioned in Section 2, pose estimation is essentially the P*n*P problem. P3P is a common method for solving P*n*P, which involves the calculation of *cosine* values for a few angles formed by the projection center of the camera and three selected 3D points and several non-polynomial operations (*e.g.,* square root and cube root). Since the *cosine* function and the required non-polynomial operations are not supported by FHE, one possible solution is to resort to alternative approaches such as direct linear transformation (DLT) [15] for solving P*n*P.

Preliminary Results. To better understand the issues of integrating FHE into spatial mapping and localization, we conduct a pilot study using the SEAL [23] homomorphic encryption library released by Microsoft. It includes two different schemes BFV [9] and CKKS [4]. The BFV scheme allows modular arithmetic operations on encrypted integers and is the only choice when exact values are required. While the CKKS scheme, on the other hand, supports addition and multiplication operations on encrypted real numbers, it generates only approximate results. Since the features include real numbers, we have to choose the CKKS scheme.

We first evaluate the computation overhead and the inflation of encrypted data caused by FHE. When calculating the pair-wise distance between localization features and those in the spatial map, it takes 18 ms on average with FHE-encrypted data. In comparison, it takes only 0.0076 ms without FHE. The size of a feature descriptor extracted by SIFT is 128 bytes. In contrast, the size of an encrypted feature using the CKKS scheme is 393,321 bytes. The above results show that it is not practical to encrypt feature descriptors with FHE. We next evaluate the localization error introduced by estimating the 6DoF pose using DLT, instead of P3P. We apply RANSAC to remove the outliers for both approaches. Figure 6 shows the normalized localization errors of DLT using the results of P3P as the baseline. Compared to P3P, for 50% of cases, the normalized localization error of DLT is higher than 3.8 for position and orientation, respectively.

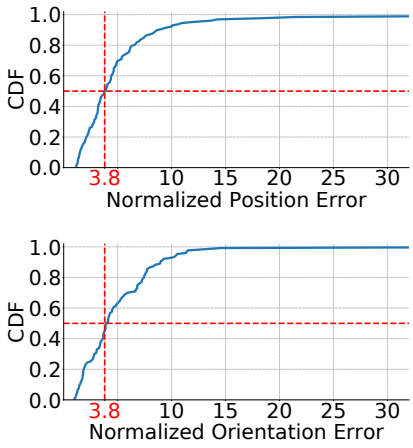

Figure 6: CDF of normalized localization errors when solving P*n*P using DLT, compared to P3P.

Discussion. Our preliminary results demonstrate that there are still several open issues when utilizing FHE for spatial mapping and localization. First, FHE cannot support accurate 6DoF pose estimation methods such as P*n*P with RANSAC. The alternative solutions that can operate on FHE-encrypted data lead to high localization errors, which demands more investigation. Second, the computation latency of operations on FHE-protected data is around 3000× higher than that of plain-text data. We will explore the applicability of recent acceleration proposals such as Pegasus [19] for spatial mapping and localization. Third, FHE results in the size inflation of encrypted data and thus increases the required network bandwidth and storage overhead, which calls for attention from the cryptography community.

4 CONCLUSION

In this paper, we proposed a holistic research agenda to preserve privacy in the emerging mobile spatial computing paradigm, by protecting sensitive information in both spatial maps and localization images. The presented research leverages principled interdisciplinary approaches that benefit from deep learning, distributed computing, and cryptography, and demonstrates the synergy of those areas in privacy-preserving mobile spatial computing. We hope our work can shed light on more research efforts to further investigate this critical problem.

5 ACKNOWLEDGMENT

We appreciate the constructive comments from the reviewers. This work was supported in part by the NSF grant CNS-2007153 and the Commonwealth Cyber Initiative, an investment in the advancement of cyber R&D, innovation, and workforce development. For more information about CCI, visit www.cyberinitiative.org.

REFERENCES

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building Rome in a Day. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2009.

[2] F. Cangialosi, N. Agarwal, V. Arun, J. Jiang, S. Narayana, A. Sarwate, and R. Netravali. Privid: Practical, Privacy-Preserving Video Analytics Queries. In *Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2022.

[3] N. Carlini and D. Wagner. Towards Evaluating the Robustness of Neural Networks. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2017.

[4] J. H. Cheon, A. Kim, M. Kim, and Y. Song. Homomorphic Encryption for Arithmetic of Approximate Numbers. In *Porceedings of International Conference on the Theory and Application of Cryptology and Information Security*, 2017.

[5] A. Dosovitskiy and T. Brox. Inverting Visual Representations with Convolutional Networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] C. Dwork. Differential Privacy: A Survey of Results. In *Proceedings of International Conference on Theory and Applications of Models of Computation (TAMC)*, 2008.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of Theory of Cryptography Conference (TCC)*, 2006.

[8] Facebook. Project Aria. https://about.facebook.com/realitylabs/projectaria/, 2020. [accessed on 05-May-2022].

[9] J. Fan and F. Vercauteren. Somewhat Practical Fully Homomorphic Encryption. Cryptology ePrint Archive, Report 2012/144, https://ia.cr/2012/144, 2012. [accessed on 05-May-2022].

[10] M. A. Fischler and R. C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[11] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha. Visual Simultaneous Localization and Mapping: A Survey. *Artificial Intelligence Review*, 43:55–81, 2015.

[12] C. Gentry. *A Fully Homomorphic Encryption Scheme.* PhD thesis, 2009.

[13] S. Greenwold. Spatial computing. Master's thesis, Massachusetts Institute of Technology, 2003.

[14] R. Hartley, J. Trumpf, Y. Dai, and H. Li. Rotation Averaging. *International Journal of Computer Vision*, 103:267–305, 2013.

[15] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2004.

[16] S. Jana, D. Molnar, A. Moshchuk, A. Dunn, B. Livshits, H. J. Wang, and E. Ofek. Enabling Fine-Grained Permissions for Augmented Reality Applications with Recognizers. In *Proceedings of USENIX Security Symposium*, 2013.

[17] K. Lebeck, K. Ruth, T. Kohno, and F. Roesner. Securing Augmented Reality Output. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2017.

[18] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Journal of Computer Vision*, 60(2):91–110, 2004.

[19] W. Lu, Z. Huang, C. Hong, Y. Ma, and F. Qu. Pegasus: Bridging Polynomial and Non-polynomial Evaluations in Homomorphic Encryption. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2021.

[20] Magic Leap. Spatial Computing: An Overview for Our Techie Friends. https://www.magicleap.com/en-us/privacy/spatial-mapping-overview-and-detail-options, 2022. [accessed on 05-May-2022].

[21] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] Microsoft. A new era of spatial computing brings fresh challenges—and solutions—to VR. https://tinyurl.com/4vzht8rp, 2019. [accessed on 05-May-2022].

[23] Microsoft. SEAL homomorphic encryption library (release 3.7). https://github.com/Microsoft/SEAL, 2021. [accessed on 05-May-2022].

[24] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 2016.

[25] F. Pittaluga, S. J. Koppal, S. B. Kang, and S. N. Sinha. Revealing Scenes by Inverting Structure from Motion Reconstructions. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[26] R. Poddar, G. Ananthanarayanan, S. Setty, S. Volos, and R. A. Popa. Visor: Privacy-Preserving Video Analytics as a Cloud Service. In *Proceedings of USENIX Security Symposium (USENIX Security)*, 2020.

[27] F. Roesner, T. Kohno, and D. Molnar. Security and Privacy for Augmented Reality Systems. *Communications of the ACM*, 57(4):88–96, 2014.

[28] E. E. Sabelman and R. Lam. The Real-Life Dangers of Augmented Reality. *IEEE Spectrum*, 52(7):48–53, 2015.

[29] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] T. Sattler, B. Leibe, and L. Kobbelt. Fast Image-Based Localization using Direct 2D-to-3D Matching. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2011.

[31] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image Retrieval for Image-Based Localization Revisited. In *Proceedings of British Machine Vision Conference (BMCV)*, 2012.

[33] J. L. Schönberger and J.-M. Frahm. Structure-From-Motion Revisited. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[34] S. Shekhar, S. K. Feiner, and W. G. Aref. Spatial Computing. *Communications of the ACM*, 59(1):72–81, 2016.

[35] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks . In *Proceedings of International Conference on Learning Representations (ICLR)*, 2014.

[36] B. Triggs. Perspective Projection. https://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MOHR_TRIGGS/node9.html, 1998. [accessed on 05-May-2022].

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[38] H. Wu, X. Tian, M. Li, Y. Liu, G. Ananthanarayanan, F. Xu, and S. Zhong. PECAM: Privacy-enhanced Video Streaming and Analytics via Securely-Reversible Transformation. In *Proceedings of ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2021.

[39] H. Wu and J. Zhou. Privacy Leakage of SIFT Features via Deep Generative Model Based Image Reconstruction. *IEEE Transactions on Information Forensics and Security*, 16:2973–2985, 2021.

[40] C. Xiang, C. R. Qi, and B. Li. Generating 3D Adversarial Point Clouds. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] K. Yousif, A. Bab-Hadiashar, and R. Hoseinnezhad. An Overview to Visual Odometry and Visual SLAM: Applications to Mobile Robotics. *Intelligent Industrial Systems*, 1:289–311, 2015.