# Inherently Privacy-Preserving Vision for Trustworthy Autonomous Systems: Needs and Solutions

Adam K. Taras[a], Niko Suenderhauf[b], Peter Corke[b], Donald G. Dansereau[a]

[a] *Australian Centre For Robotics, School of Aerospace, Mechanical and Mechatronic Engineering, The University of Sydney, 2006, NSW, Australia*
[b] *Queensland University of Technology (QUT), Centre for Robotics, Brisbane, 4001, QLD, Australia*

**Abstract**

Vision is an effective sensor for robotics from which we can derive rich information about the environment: the geometry and semantics of the scene, as well as the age, identity, and activity of humans within that scene. This raises important questions about the reach, lifespan, and misuse of this information. This paper is a call to action to consider privacy in robotic vision. We propose a specific form of inherent privacy preservation in which no images are captured or could be reconstructed by an attacker, even with full remote access. We present a set of principles by which such systems could be designed, employing data-destroying operations and obfuscation in the optical and analogue domains. These cameras *never* see a full scene. Our localisation case study demonstrates in simulation four implementations that all fulfil this task. The design space of such systems is vast despite the constraints of optical-analogue processing. We hope to inspire future works that expand the range of applications open to sighted robotic systems.

*Keywords:* privacy-preserving vision, optical computing, robotic imaging, localisation

## 1. Introduction

Do you have a robot vacuum cleaner? Perhaps one of the new generation robots that uses a camera to navigate around your house? Where do those

*Email addresses:* adam.taras@sydney.edu.au (Adam K. Taras), donald.dansereau@sydney.edu.au (Donald G. Dansereau)

camera images go? Who can see them? Perhaps the images should never leave your house. Perhaps they should never leave the robot or the camera chip. Perhaps, to best protect your privacy, the images, as we know them, should never be formed in the first place.

Research in robotic vision has neglected and often ignored the legitimate privacy concerns of potential end-users, and instead focused solely on improving task performance [1]. In one high-profile example [2] private images from a robot vacuum cleaner found their way to social media via a data labeling provider. We argue that this current disregard of privacy ultimately forestalls the widespread adoption and societal impact of robotic vision. In contrast, we propose to re-imagine robotic vision to achieve an optimal balance between task performance and privacy protection.

This paper is a call to action for the robotic vision community to develop novel computational imaging technology for inherently privacy-preserving robotic vision. By developing novel combinations of optical, analogue, and algorithmic elements, the community – academia and industry – could build novel camera technology that never forms human-interpretable conventional images, and from which such images could never be reconstructed from the sensor data.

Such new camera designs would address the legitimate privacy concerns that are impeding the beneficial adoption of robotics in applications of societal and economic importance, e.g. where there is a strong emphasis on social human-robot interactions (healthcare, aged care); where robots and humans collaborate and intellectual property must be protected (manufacturing); or where a breach of privacy could have safety and security implications (energy) or impede sovereign capabilities (defence). By addressing these legitimate privacy concerns, novel privacy-preserving camera technology will broaden the applicability, and increase the public acceptance, of robotic vision applied to these domains without compromising the privacy and security of citizens, industries and governments.

In this paper we introduce privacy as a concept in the context of robotic vision (§1.1) and discuss current approaches to privacy preserving robotic vision (§2). We then introduce our proposed concept for inherently privacy-preserving vision systems (§3) and present a localisation case study that exemplifies this approach with four different implementations (§4). We close with a discussion (§5) outlining our call to action for a concerted effort by the community to generalise these concepts beyond localisation and consider inherently privacy-preserving vision as a challenging yet extremely valuable

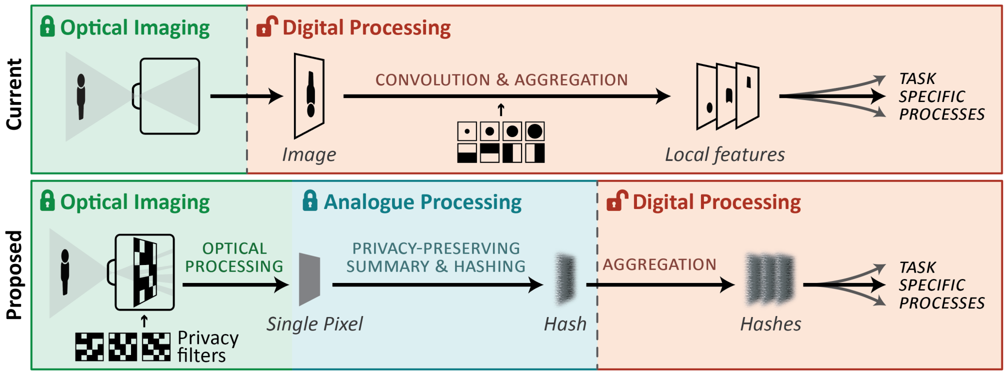

Figure 1: Example of the proposed inherently privacy-preserving approach to vision: Current robotic vision (top) uses conventional optics and cameras to form human-interpretable images with many 100,000s of pixels. Local features are often extracted through a series of learned convolutions and aggregations, e.g. by convolutional neural networks, before being processed by task-specific components, e.g. for object detection or grasp synthesis. Instead, we propose to shift processing into the optical-analogue domain (bottom): in this example a micromirror device implements a series of learned filters through which the light in a scene enters a single-pixel sensor. Analogue processing performs privacy-preserving summarisation and hashing *before* the data enters the digital domain, where it is vulnerable to attacks and security breaches. Specialised task-specific algorithms and learned components operate on the secure hashes in the digital domain to perform important robotics tasks. Private data can *never* be directly captured by the system, and the digital data is obfuscated by the architecture such that reverting the optical-analogue hashing is intractable. We show that such a system is capable of localisation, and believe this could be generalised to other tasks.

ambition.

The project page https://roboticimaging.org/Projects/Privacy/ has code and datasets. As part of the call to action, we also encourage the vision community to attack the privacy preserving system we propose and attempt to solve our reconstruction challenge, see website for details.

### *1.1. What is Privacy?*

Privacy is a complex concept that is relevant to many areas of society. Interestingly, it was the increased availability of easy-to-use photography cameras that motivated the definition of privacy as “the right to be let alone” in an 1890 law review article [3].

Since then, numerous definitions and analyses of privacy have been published, with Altman’s “selective control of access to the self or to one’s group" [4] one of the most prominent. As reviewed in [5], the current research literature distinguishes physical, psychological, social, and informational privacy. These respectively relate to concepts such as personal space or physical access; the right to control with whom and under what circumstances to share one's thoughts; the ability to control anonymity and social interactions; and when, how and to what extent information about the self will be released to another person or organisation [5].

Although all forms of privacy can potentially be violated by robots and are therefore relevant to the study of robotics, our project focuses on a specific form of informational privacy. Concretely, in this paper, we understand *privacy preservation* to be the minimisation of the risk of exposing a human-interpretable image of the environment in which a robotic vision system operates, or the risk of exposing information that enables the reconstruction of such an image.

The robotics community largely considers privacy concerns and task performance to be orthogonal issues: of 89,120 papers published in the top robotics journals and conferences 1982-2019, only 0.5% mention privacy [1], despite the fact that at least 132 countries now have data privacy laws, and data protection officials from over 60 countries have expressed concerns about the impacts of robotics and AI on privacy [1]. The disconnect between the robotics research community and these recent developments is a clear call to action, and our motivation.

Beyond the robotics community, other fields are similarly grappling with this challenge and broadly agree it requires attention. A recent survey of internet of things literature [6] found "there is a lack of efficient privacy and security algorithms for IoT", as well as a recent outlook in machine learning finding [7] "[works] on the preservation of privacy... are still in an infancy stage". Our contributions demonstrate that ambitious architectures can enable inherent privacy, and that our approach of never capturing privacy-revealing data should serve as best practice.

## 2. Current Approaches to Privacy in Vision

Much of the current work for privacy preserving vision seeks greater levels of obfuscation or even encryption. These works fall broadly under two categories, depending on when the obscuration occurs: after capturing digital images, or during the image formation process. Here by obfuscation we mean representing signals in a difficult-to-interpret form, and by encryption we entail encoding with a cryptographic key such that information can only be compromised if someone knows or deduces the key.

In this section, we discuss encouraging results spanning these approaches, and identify key gaps that represent the opportunity at the heart of this paper.

Even among those tackling the specific task of privacy-preserving vision, there are multiple definitions of privacy in use. For example, Zhang et al. [8] propose the use of red green blue-depth (RGB-D) cameras for masking foreground objects and using only depth information to maintain user privacy. Many would also argue that capturing of depth-only imagery is an unacceptable breach of privacy. In our work, privacy-preserving means that at no point in the system are digital images stored, nor could they be reconstructed.

The fundamental law of information recovery [9] states that "overly accurate answers to too many questions will destroy privacy in a spectacular way". We use this as a guiding principle in understanding current approaches to privacy preservation, employing the most conservative position that anything that is digitised should be suspect as open to unwelcome observers. While vision systems may summarise images with sparse representations, e.g. through feature extraction, Pittaluga et al. [10] show that even sparse information like 3-dimensional (3D) point clouds of scale invariant feature transform (SIFT) features can be used to reconstruct colour images of the complete scene. This reconstruction succeeds even after removing keypoint orientation and scale.

Reconstruction from features is a strong cautionary result. Even if feature extraction is implemented before digitisation it is not guaranteed to be privacy preserving, but rather the features themselves must be robust against reconstruction and various attacks. Thus, key characteristics used to judge the privacy-preserving capabilities of systems are the amount of information captured as well as by what means information is either discarded, obfuscated, or encrypted.

### *2.1. Post-Capture Obfuscation and Encryption*

Methods in this category capture digital images then extract and obscure key information before discarding the images. Any system that captures digital images is open to attack via unsecured remote access to the compute system. Of even greater concern is that in some cases the obscured form of the images that are generally taken to be privacy-preserving nonetheless provide enough information to reconstruct imagery of the scene. A concrete example of the later arises in the random feature lines proposed by Speciale et al. [11, 12]. This work summarises scenes in terms of obfuscated keypoints, replacing SIFT features with randomly oriented lines in three dimensions. This obfuscates the contents of the feature cloud, rendering scenes unrecognisable to the human eye. However, Chelani et al. [13] later found that images can be reconstructed from these obscured feature clouds by exploiting statistics of nearby feature points. This cautionary example indicates that there is a wealth of structure and redundancy in even obscured representations, and points to the need for a more significant re-think of how we carry out vision to preserve privacy.

Conventional encryption [14], whilst relatively inexpensive and broadly applicable, is also subject to breaches. Encrypted data may be open to unauthorised access, and data breaches due to human error are uncomfortably common and increasing in frequency [15]. Ultimately, encrypted data stored in the cloud [16] is only as secure as the agents entrusted with the decryption keys. This opens the potential for spoofing attacks, requires securely distributing keys, or requires algorithms that work on encrypted imagery without requiring decryption, all of which are challenging.

### *2.2. Optical Obfuscation and Encryption*

There are several notable examples where the camera is involved in the process of obfuscating or encrypting imagery. Key-Nets [17] use custom optical fibre bundles and custom imaging sensors with per-pixel bias and gain to effectively carry out a Hill Cipher [18]. The paper proposes a way of converting neural architectures to operate on the encoded imagery, offering the potential for existing vision pipelines to be used on keyed data. However, this approach requires custom optical fibre bundles to shuffle pixels and custom silicon implementing per-pixel bias and gain. Manufacture is thus extremely impractical, preventing widespread adoption of this approach. It is also the case that this approach, like other forms of cryptography, is only as secure as storage of a private key, and is open to a variety of attacks.

There is growing interest in the use of optical neural networks including diffractive deep neural networks [19, 20]. Recent work shows that diffractive layers can be designed such that destructive intereference occurs for all except a target class. The camera captures data with low latency as all computation is optical, and manufacture of printed masks is practical and broadly accessible. A key limitation is that the proposed diffractive cameras require narrowband, collimated active illumination to function, limiting their widespread deployment. Hard guarantees about leaking of private information through the diffractive imaging process are also unclear.

In addition, Horton et al. [21] demonstrate task-specific operations directly on bytes generated by a very simple privacy preserving camera that only uses a fraction of the pixels. It is unclear if such a method is robust to reconstruction attempts, which could succeed since the proposed pixel subset is fixed and could be solved for by an attacker. We complement this by providing details of how to process the incoming image in an inherently privacy-preserving and obfuscating way.

Finally, there is growing interest in reconstruction-free vision systems, in which encoded imagery is captured but never converted to a human-interpretable form. An example of this is the use of a lensless imaging for action recognition [22]. Whilst capturing obfuscated imagery, this work does not explicitly address the potential for an attacker to reconstruct human-interpretable imagery. Based on recent progress in lensless imaging that explicitly reconstructs human-interpretable imagery [23], it seems inevitable that some form of reconstruction should be possible in these cameras. Another example of these systems is image classification from single-pixel cameras [24]. Such systems are likely vulnerable to reconstruction as evidenced by techniques in compressive sensing.

### *2.3. Key Gaps*

It is evident that prior works either digitise signals and then make them private, and are open to digital attack; or involve the camera and do not prevent reconstruction. A single exception, Key-nets, is impractical due to the requirement for custom optical and silicon manufacture.

In the following we propose an approach to inherently privacy-preserving vision that involves imaging the scene through optical-analogue computation such that only secure, privacy-preserving information is ever transferred into the digital domain. Relative to prior work, key differences are in the nature and quantity of digital information, and that reconstruction should not be possible. We leave applying inversion attacks on the methods presented as future work, however the results of the case study compel us by illustrating how sparse the data is and how ill-posed the inversion problem is.

Recent work [25] asserts that *any* localisation algorithm that produces poses can be attacked through queries of database images sourced from e.g. the internet. Even if our framework was modified to produce poses instead of matches along a previously traversed route, we believe our method is resilient to these attacks due to the use of global features. To invert a global feature extractor, an attacker would have to combine multiple objects, from which the number of possible combinations grows intractably large. This family of attacks could also be mitigated by taking data that is not reliably found in large quantities easily, for example by capturing non-anthropomorphic data. This warrants further work.

Our approach is extensible so that depending on application, other definitions of privacy can be engineered into the camera. For example, in some cases making an inference of the existence of an object in the scene could be a violation of privacy. This could be approached in an adverserial setup, where the feature extractor aims to confuse a binary classifier. Thus, we believe this represents a unique approach to privacy preservation, opening new application areas where systems with vision cannot operate at present.

By judiciously moving computation out of the digital domain, removing and summarising information, and obfuscating or encrypting the information that remains in the optical-analogue domain, our proposed approach offers a different class of privacy-preserving camera that we call inherently private vision systems.

## 3. Inherently Privacy-Preserving Vision

In this work we propose the concept of an inherently privacy-preserving vision system. This is one in which none of the attacks outlined in the previous section could be applied: brute-force decryption, image decoding, spoofing attacks, data breaches, and access to digital imagery through unauthorised remote access to a robot's hardware should not be possible. Privacy in this sense means that at no point in the system are digital images stored, nor could they be reconstructed.

Inspired by work on custom optics and sensors [17] as well as optical neural networks [20], we propose that inherently private vision is possible by constructing custom cameras designed to carry out specific tasks chiefly through their optics and analogue electronics. These cameras must be designed following a set of principles that prevent the digitisation of private information.

We propose here a starting point for this set of ideals for constructing inherently privacy-preserving vision systems:

- Specialise the camera to the task; this sacrifices generality for privacy as the camera can only be used for the task(s) it's designed for,

- Shift as much processing as possible out of the digital domain, keeping it out of reach of remote attack,
- Maximise information-destroying operations prior to digitisation,
- Apply obscuration prior to digitisation such that brute-force attack becomes the only option for inverting the imaging process,
- Consider all information already available to the attacker, e.g. sequences of data and priors to improve domain performance and ensure privacy, and
- Maximise ambiguity, so that even a successful brute-force inversion of the imaging process is not likely to yield the correct image.

Such sensors could fit into existing systems relatively easily. The software reading from the sensor could still be updated, however the secure, optical / analogue processing must remain offline and therefore beyond the reach of software updates and attacks. Our approach thus produces cameras that trade flexibility for privacy.

We anticipate a broad variety of implementations could meet the above principles, and in this paper we conduct a case study for carrying out localisation with only four of such methods. Comparison of conventional imaging and our specific implementation employing optical-analogue single-pixel hashing are depicted in Figure 1, and a detailed proof-of-concept study carried out in simulation is included in the following section. In this implementation we address the principles laid out above by specialising a camera to the task of localisation, shifting much of the machinery of localisation into the optics and analogue electronics of the camera, and employing information-destroying feature extraction, summarisation, and hashing prior to digitisation such than many images yield identical hashes that are nevertheless useful for localisation.

## 4. Case Study: Privacy-Preserving Localisation

Here we apply the principles laid out in the previous section for designing inherently privacy-preserving vision systems for the robotic task of localisation. We show in simulation that by shifting digital processing into optics and analogue electronics, we can accomplish effective localisation without ever capturing digital images and without capturing enough information to allow reconstruction of these images.

The overall approach of this study is to build on the architecture of a single-pixel camera to carry out feature extraction and summarisation in the optical-analogue domain. In this architecture a series of masks are applied to a wide-field single pixel, and the resulting signal passed through bespoke analogue computation. Applying insights from existing feature-based methods, we select masks and subsequent summarisation that represents a sort of hashing or fingerprinting, such that information is destroyed before digitisation. We evaluate the proposed approach by solving an image retrieval problem analogous to localisation, demonstrating accuracy on par with a standard SIFT-based approach.

### *4.1. Why Localisation?*

A robotic vacuum cleaner working in medical settings, a warehouse cart handling sensitive intellectual property in manufacturing and a drone delivering goods over government buildings all rely on an understating of position within the scene. With current vision systems, however, any data collected would be at risk to attack. By addressing the problem of privacy-preserving localisation, we address privacy issues for a breadth of application domains. We anticipate that solving the localisation problem can also give direct insight into more complex vision tasks such as object tracking or grasping.

Localisation tasks range from image retrieval [26, 27], place recognition [28], or full 6 degree of freedom pose regression [29, 30]. Here we address the image retrieval problem for privacy preserving localisation.

### *4.2. Optical-Analogue Image Hashing*

In this section we develop some concrete implementations of inherently privacy preserving architectures. We will present four methods of processing the images in the optical domain and one method of processing the signals in the analogue domain, indicating that there is still a large space of designs possible within the framework outlined in Figure 1.

We first identify and design methods of image retrieval such that the signal processing in the optical-analogue domain computes image hashes with high utility but limited private information. The hash function must be tractable in hardware, information-destroying, and descriptive enough to allow localisation. We note intuitively that local hashes may be more open to exploitation for reconstruction as they reveal local structure, where an attacker needs to solve multiple smaller inverse problems. This intuition is supported by recent work [25]. Global hashes, on the other hand, couple

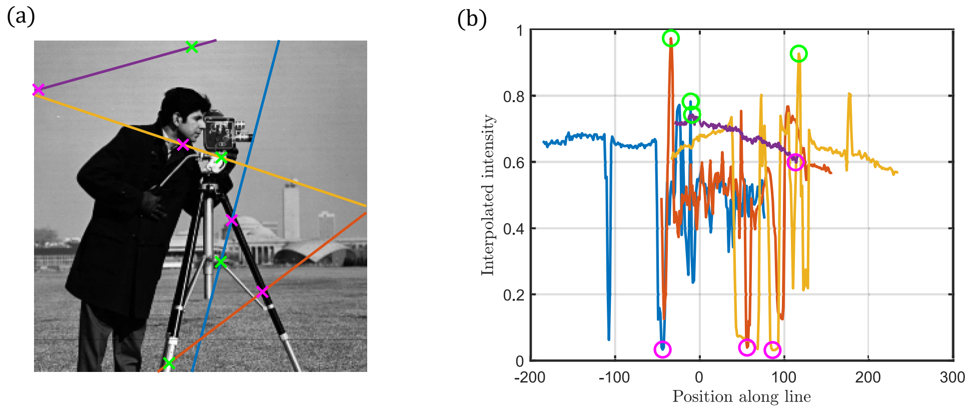

Figure 2: Optical-analogue summarisation for localisation: (a) an input scene is sampled along four randomly generated lines, recording only the global maxima (green) and minima (magenta) along each line, (b) traces along each line with extrema highlighted. The hashing process accumulates these extrema prior to digitisation, destroying and obscuring information about the scene.

information between different structures, introducing more ambiguity. Thus, we restrict our pipeline to using global features.

Edges play an important role in image understanding, and in the context of our single-pixel architecture a simple way of looking for edges is to admit light along lines. This motivates our first proposed implementation.

*Random Lines.* We measure extrema along the masked lines by using maximum and minimum hold circuits. We accumulate the resulting pairs of maxima and minima, one per line, over $N$ lines. For randomly selected and ordered lines, the resulting accumulation of pairs of extrema reveals little about the structure of the image, while representing a fingerprint that can be used to discern the image from a sequence.

We illustrate the process of measuring this hash for $N = 4$ features in Figure 2, where each feature is the tuple of maximum and minimum along the curve. To look for more edges we measure along more lines, and by randomising the locations of these lines we destroy information about the original structure of the scene while collecting a fingerprint of its content. Note that, by drawing from uniform distributions, the resulting collection of features is rotation-invariant for large $N$.

We depict a hardware implementation of the line hash in Figure 3. To maintain privacy the DMD is driven through a fixed set patterns that cannot

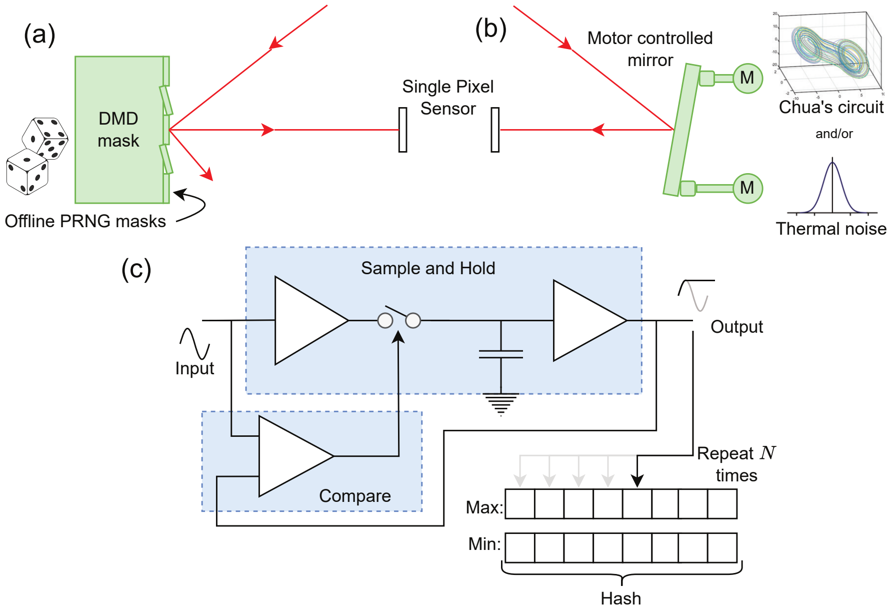

Figure 3: Hardware implementation for processing in the optical (green) and analogue (blue) domains. We compare two methods for the optical domain. A DMD controlled by an offline pseudo-random number generator (PRNG) (a) applies a set of fixed filters to light from the scene, or a controlled mirror (b) places parts of the image over a single pixel sensor using chaotic dynamics from Chua's circuit and/or thermal noise. The signal from the sensor traverses analogue computation (c) that detects extrema over the series of filters. This is repeated over $N$ features, accumulating a histogram or hash-like fingerprint of the scene that is then digitised and loaded to perform localisation. See Figure 4 for example hashes.

be changed. This limits flexibility of the camera, but is critical to prevent outside attack.

*Random Circles.* For the next presented method, we alter the masks employed by the DMD. One concern with the random lines is that they begin and end at the boundaries of the images, and under motion this could allow an attacker to infer details of the image boundaries. This motivates an alternative hash which computes extrema over circles rather than lines. Selecting random radii and positions yields similar properties to the random-line approach, including rotation invariance.

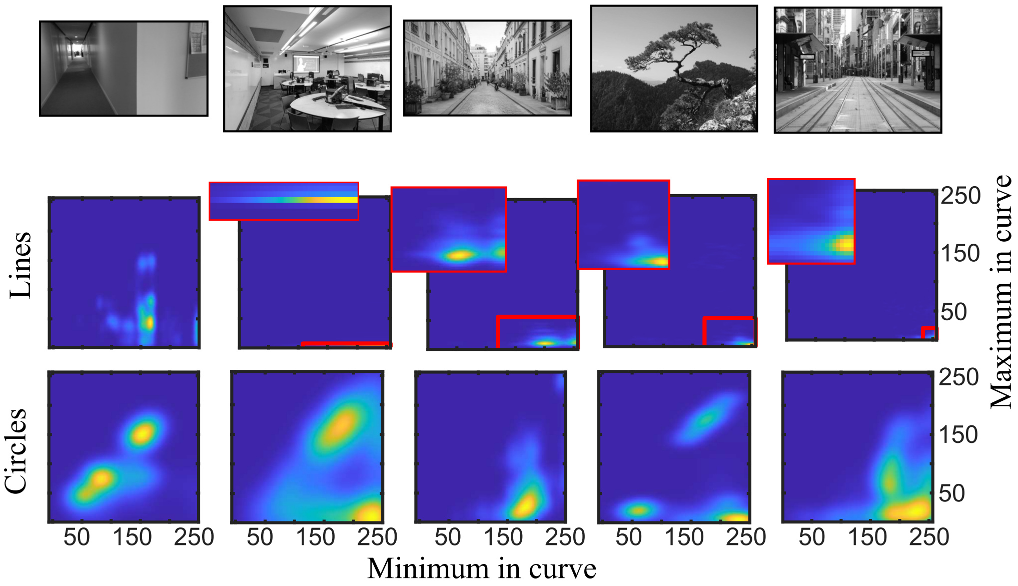

Figure 4: Comparison of global feature fingerprints for different scene types, with each column showing a different location. $10^3$ random line extrema features (middle row) and random circle extrema features (bottom row) with radii sampled from $[15, 30]$ pixels show that each image has a unique fingerprint. Insets depict greater detail around features with saturation.

*Analogue Randomness: Thermal Noise and Chua's Circuit.* A further iteration on the above addresses the weakness of using a PRNG to select masks. If an attacker was to obtain the seed or state of the offline system, all future masks are deterministic. We present two possible methods of removing this dependency – using thermal noise or Chua's circuit to provide the analogue signals to control a mirror, thus positioning a controllable patch of the image on the single pixel sensor. The motor positions are determined by a superposition of two analogue signals – a sample and hold value for the centre of the feature circle and a oscillator to trace out the circle around it. The radius of the circle is controlled by another held analogue signal. The details of the simulation method and parameters are included in Appendix A.2. These implementations further increase the difficulty of a remote attacker reasoning about the optical computing.

Figure 4 visualises the proposed hashes by plotting histograms of extrema pairs in 2D. These are shown for a range of input scenes. Here we compute hashes over $N = 10^3$ random lines or circles, and smooth the display for visualisation using kernel density estimation [31, 32]. By construction, this visualisation of the hashes must lie below the diagonal. For the random line extrema, each line measures the dominant edge over the dimension of the image. Since this spans the whole image, the features present are strongly affected by the amount of saturation in the image. On the other hand, the random circle extrema are more sensitive to local edges and features. When these features lie close to the diagonal, the scene has textureless regions. In the following we show that these hashes represent fingerprints that are sufficiently unique to allow localisation.

### *4.3. Localisation*

To localise based on the proposed hashes, we train a bag of words (BoW)-based approach on a dense trajectory of hashes. At inference time, a query hash from the sensor is presented for search and the most similar hash in the reference trajectory is retrieved, localising the robot to the corresponding point in the trajectory. This approach supports a variety of types of visual words, allowing us to directly compare our hashing approach with more conventional, privacy revealing features. We choose BoW over neural network-based approaches because, although they may offer superior results, they also show more complex behaviours that can be more difficult to interpret. That our approach works well even with the simpler BoW localisation is more revealing.

### *4.4. Results*

We evaluate how well BoW-based image retrieval is able to predict the position of unseen test images. We use the “Digiteo Seq 2” dataset [33], which contains handheld photos from an office. We use a single camera for the trajectory from this stereo dataset. We also have produced our own indoor datasets, which will be made available. These contain colour images from a mobile phone on a gimbal mount. There are two subsets: PNR, which has two videos of the same room, and ABS, which has four trajectories of 130 images: normal, rotated (10-15°), translated (15-30cm), and rotated + translated, with all image indices matched correctly.

The BoW is trained on hashes from a training stride of one in every 20 images, and the remaining images are used for testing. When querying a test image in the same dataset, we consider the localisation to be correct if the true image index is within 30 frames (i.e. within one second of motion, or a fraction of a metre for a rolling platform travelling at a typical speed) of the BoW best-match image.

In Figure 5, we measure the accuracy of localisation while varying the number of curves used in our approach and compare against the same BoW approach trained on SIFT features. For few features performance is weak, but it increases as the number of features is increased. We also consider randomizing the feature curves, circles or lines, for each input image, or using a fixed set. The difference between these is not significant, and it decreases as the number of features increase.

The proposed methods are ultimately able to meet and slightly exceed the baseline performance of SIFT, with no significant advantage to either lines or circles as curves in this case. The slight variation in SIFT performance indicates that fine tuning of the image retrieval system is useful in optimising performance for a particular dataset or context. We do not claim that these methods outperform SIFT, but rather that they are comparable in localisation accuracy while inherently maintaining privacy.

Empirically we observe that $N \sim 10^3$ curves are required for good performance at this task. This is a factor of $10^3$ fewer than the number of pixels in the input (megapixel) images. We investigate which parts of the input images are reflected in the proposed hashes in Figure 6. We see that the hashing procedure both reduces the amount of image information represented and obscures it in keeping with the recommendations in the previous section. The hashing only reveals a small subset of the image, the hashing process hides the locations of the extrema, and the distributions of maxima and minima do not directly reflect the intensity distributions present in the images.

From these observations we conclude that it is doubtful any algorithm could reconstruct an image from its hash. However, even if one did succeed at this, the fact that a very large number of distinct images produce the same hash would prevent the attacker from knowing if they had constructed the correct image.

We also experiment with how to make the features robust to variations in brightness, and sensitivity to spacing in the training data. Figure 7 shows the results of our experiments. In (a), we see the method is relatively insensitive to training stride. We also consider the use of a contrast metric $\max - \min$ over each feature to weigh the importance, and observe this has little effect. A study of other possible metrics is left as future work. In (b) we show that normalisation improves robustness to brightness variation. This can

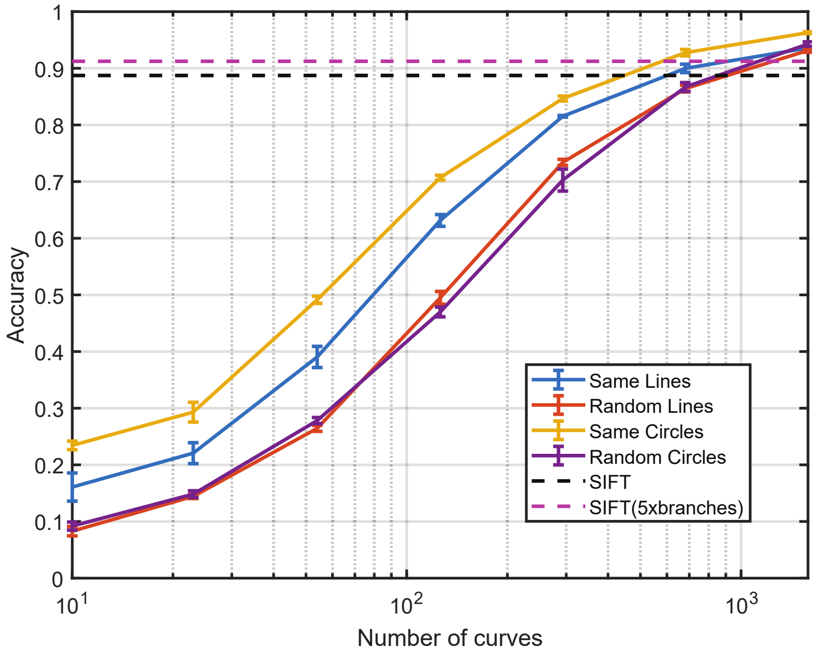

Figure 5: Accuracy of localisation as a function of the number of features accumulated $N$, and for randomly changing curves or repetition of the same curves for each input scene. The SIFT localisation approach is also shown for comparison. While initially performing worse, for large $N$ using randomly changing curves converges to the performance of fixed curves. There is no significant difference between the circle and line methods, and performance exceeds that of the conventional SIFT-based approach.

be seen in performance improvement at lower brightness factors. However, there is an unexpected drop in accuracy at all brightness factors. We believe this indicates there is useful information in absolute brightness. Thus, a more robust privacy-preserving feature extractor would use both absolute brightness and normalised values. This is a topic for future research.

Finally, we compare different feature extractors over many datasets. We also include oriented FAST and rotated BRIEF (ORB), a local feature extractor aimed for use on real time platforms due to lower computational cost than SIFT. Our results are presented in Table 1. In cases where the training and testing sets are the same, all results are strong, indicating that these methods can localise well for positions directly between known states. The Chua's circuit method performs best. In cases where the training and evaluation are different, SIFT is clearly the strongest, however our methods are

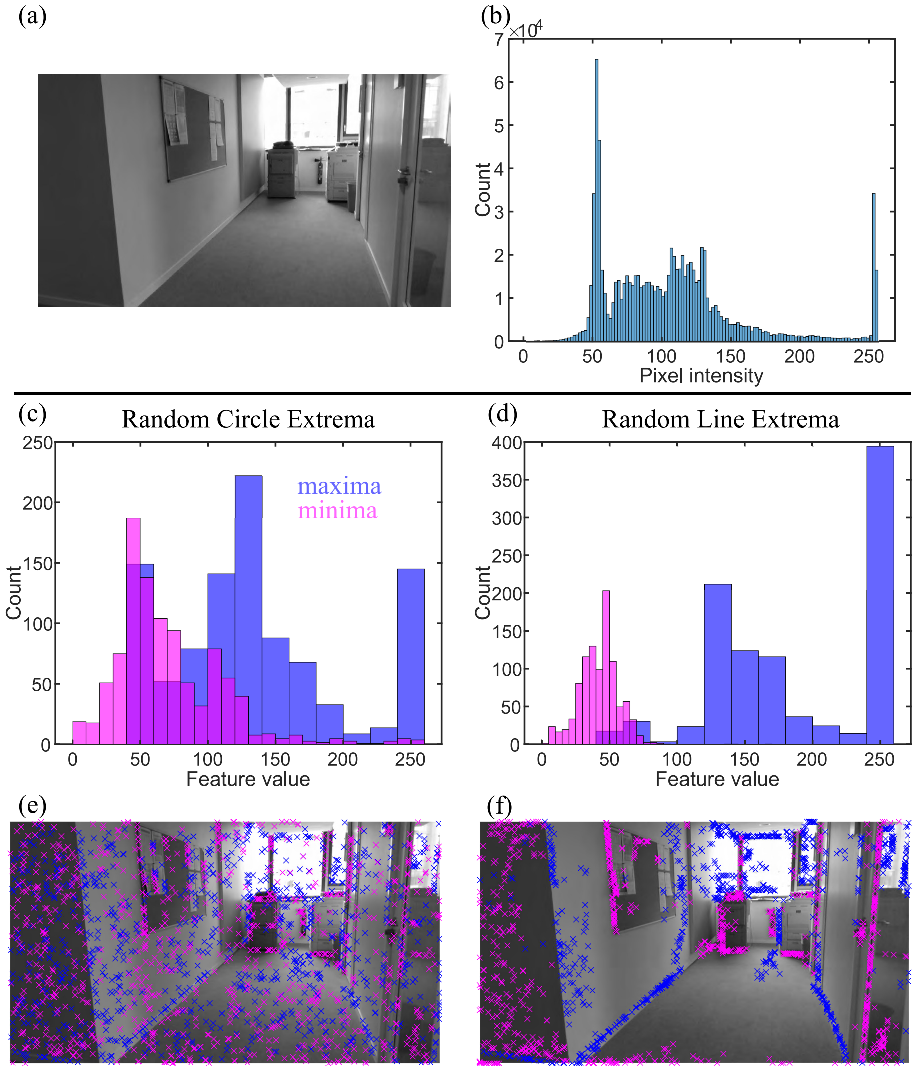

Figure 6: Distribution of sources of data for proposed hashes. (a) original scene (b) distribution of pixel intensities in the scene. Once processed by our random circle extrema pipeline (c,e) or random line extrema (d,f) the distributions do not sample the true distribution evenly. The visualisation of the source of data (e,f) indicates that there are large sparse regions in either case, with lines sampling more densely around extreme regions as expected. The true locations of the data in the hash is not exposed, so an attacker would have to recreate the image without knowing the location of the extrema, only their intensities.

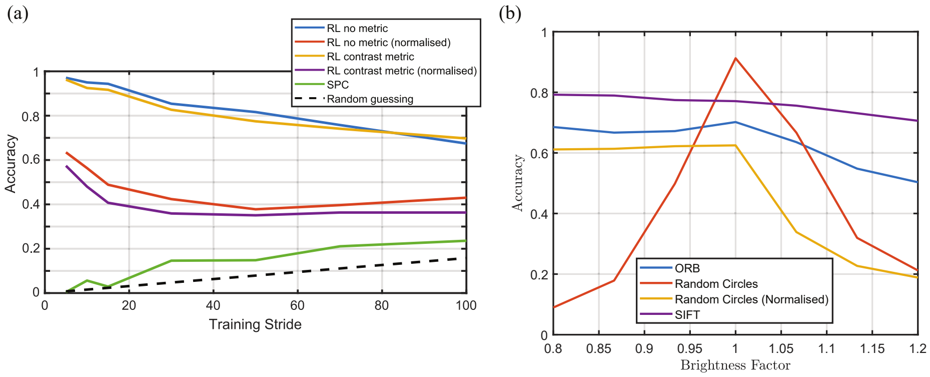

Figure 7: a) Accuracy of localisation for different training strides (number of images skipped in a trajectory to train on) comparing the random line (RL) feature extraction to a single pixel camera (SPC) that takes the mean of the image. Increasing the training stride decreases the amount of feautres the system is exposed to and increases the motion between known and tests frames, decreasing performance. Normalising the features to the mean of the image decreases performance, indicating that the mean holds discriminating power. However, the mean alone (SPC) performs poorly. b) Accuracy of localisation as a function of the brightness scaling (including saturation effects) applied to the test image, after training on a unaltered set. Absolute extrema are strongly effected by even small changes in brightness and falloff rapidly. Normalising random circle features leads to robustness for dimmer scenes but saturation results in reduced performance. Local methods produce features that are more robust to these effects.

equally or more robust that ORB. In these experiments, all methods have the same number of features. We hypothesise that the Chua's circuit method is the best performing from our implementations since it covers a smaller area of the image at a higher density. More complicated masks, however, could use the global field of view of the DMD to measure more of the image at once. Optimization of the optical-analogue processing is an exciting avenue of future work.

An outline of the hyperparameters used in all feature extaction methods can be found in Appendix A.1. We tuned these parameters manually to allow us to draw the most important generaliseable conclusions, however anticipate a more rigorous approach such as a random search over a target dataset could further improve task performance.

We also justify the use of interpolation in simulation. For the motorised mirror case, the sampling is truly continuous, while for the DMD implementation the maximum resolution of commercially available parts is up to

Table 1: Comparison of localisation accuracy (%) over different datasets. Our methods are all run with $N = 3000$ features. On high resolution datasets, the privacy preserving methods are of similar utility to conventional privacy revealing methods SIFT and ORB.

| Train Set | Test Set | SIFT | ORB | UC [1] | UL[2] | TC[3] | CC[4] |
|---|---|---|---|---|---|---|---|
| Digiteo_seq_2 | Digiteo_seq_2 | 92.02 | 92.96 | 90.30 | 89.20 | 95.46 | 95.15 |
| PNR | PNR | 98.19 | 86.28 | 91.70 | 89.17 | 90.61 | 94.95 |
| PNRrotated | PNRrotated | 86.64 | 90.97 | 95.31 | 92.78 | 97.83 | 96.75 |
| ABS_norm | ABS_rot | 84.09 | 68.18 | 65.91 | 63.64 | 63.64 | 79.55 |
| ABS_norm | ABS_trans | 90.91 | 72.73 | 70.45 | 70.45 | 70.45 | 72.73 |
| ABS_norm | ABS_rot_trans | 84.09 | 63.64 | 61.36 | 63.64 | 70.45 | 72.73 |

[1] Uniform-drawn circles, [2] Uniform-drawn lines

[3] Thermal (Gaussian)-drawn circles, [4] Chua's circuit-drawn circle centres

3840×2160 pixels (e.g. TI DLP781TE), which is a factor of 4 times larger than dataset images.

## 5. Conclusions

We proposed a new class of inherently privacy-preserving vision system that fills important gaps in current approaches and opens opportunities for follow-on work. We described a set of principles by which such systems can be designed, moving information scrambling and destroying processing out of the digital domain, and thus private data out of reach of remote attackers. The proposed systems never capture images nor do they capture enough information to allow reconstruction of private images.

We demonstrated our approach through a case study in inherently privacy-preserving localisation. The success of this study lends support to the practicality of our approach. By demonstrating the feasibility of four different implementations, we see that, despite the restrictions of optical-analogue processing, there is a vast, expressive family of feature extractors to be discovered.

More broadly, this work is a call to action for the robotic vision community. We see a path forward for establishing the trustworthiness of inherently privacy-preserving vision systems:

- Characterising and refining hardware implementations based on the approaches proposed here;

- Establishing meaningful metrics for privacy in the context of optical-analogue processing and hashing;
- End-to-end design of optical, analogue and digital processing to tackle a broader range of vision tasks;
- Establishing trustworthiness through rigorous attack and refinement of the concepts in this paper;
- Communicating and educating in an accessible manner to address the barriers to societal acceptance of sighted, privacy-preserving systems.

Privacy concerns presently prevent deployment of robotic systems in important contexts including healthcare, manufacturing and defence; rising to this challenge presents a unique opportunity to benefit an otherwise unreachable part of society.

## Appendix A. Implementation Details

### *Appendix A.1. Method Hyperparameters*

The hyperparmeters in this work were tuned manually. We found a wide range of parameters to suit the localisation of a single dataset. Fine tuning (through e.g. random search) could lead to improvements in task performance.

For the uniform circles with input images of dimension $1280 \times 720$, radii drawn from the uniform distribution of $[15, 50]$ pixels showed strong results.

The centre of the circles in the thermal noise method were generated by $(x, y) \sim \mathcal{N}(S/2, S/4)$, where $S$ is a vector of the image size. Centres that were outside the image were saturated to the boundary of the image. The factor $1/4$ for the standard deviation showed good results but could be tuned further.

For both the thermal and Chua's circuit circles, radii are drawn from a Gaussian noise source. If the drawn value is below 2 pixels, it saturates to 2 pixels. This can be implemented in analogue hardware using a operational amplifier. The distributions that gave good performance were $\mathcal{N}(60, 20)$ for the thermal implementation and $\mathcal{N}(40, 20)$ for the Chua's circuit implementation.

### *Appendix A.2. Chua's Circuit Implementation*

To generate the sampling patterns for the Chua cirucit-based features, we use the dynamics of the form [34]

$$\dot{x} = -\alpha(x - y) - \alpha f(x), \tag{A.1}$$

$$\dot{y} = z - (y - x), \tag{A.2}$$

$$\dot{z} = -\beta y - \gamma z, \tag{A.3}$$

where derivatives are with respect to a dimensionless time $\tau$, $x = x(\tau), y = y(\tau)$ as voltages over capacitors, $z = z(\tau)$ as the current through the inductor, $f(x)$ is a nonlinear function describing the Chua diode. For this implementation, we use the form

$$f(x) = m_1 x + \frac{1}{2}(m_0 - m_1)(|x + 1| - |x - 1|), \tag{A.4}$$

where $m_0, m_i$ are slope coefficients determined by resistor values. We refer the interested reader to previous work [34] for details about how to convert from the dimensionless form above to circuit parameters.

To generate our data, we use parameter values of $\alpha = 15.6,\ \beta = 30,\ \gamma = 0,\ m_0 = -15/7,\ m_1 = -5/7$. The dynamics are solved using a 4th order Runge-Kutta method [35], for 180, 000 time units with initial conditions of $(0.1, 2, 1.4)$.

This is reduced to two dimensions through a linear transformation by multiplication with the matrix

$$\big(x_c, y_c\big) = \big(x, y, z\big) \begin{pmatrix} 0.0204 & -0.0847 \\ 0.0911 & -0.9490 \\ -0.1196 & -0.0392 \end{pmatrix}, \tag{A.5}$$

where $(x_c, y_c)$ is the position of the centre of the circle. This transformation is the one that minimises the standard deviation (a proxy for maximising coverage) of the trajectory. This is then scaled linearly such that the bounds of the simulation data correspond to the bounds of the image. All of these transformation could be implemented using standard gain and addition circuits from the signals in Chua's circuit directly.

To sample the locations of the feature, cubic spline interpolation is used. For each image, the feature extractor starts at a random time in the simulation range and steps forward 3 seconds between features to generate the chaotic data used in the centre of the circles.

## References

[1] S. Eick, A. I. Antón, Enhancing privacy in robotics via judicious sensor selection, in: 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 7156–7165.

[2] E. Guo, A Roomba recorded a woman on the toilet. How did screenshots end up on Facebook?, MIT Technology Review (December 2022).

[3] L. Brandeis, S. Warren, The right to privacy, Harvard law review 4 (5) (1890) 193–220.

[4] I. Altman, Privacy A Conceptual Analysis, Environment and Behavior 8 (1) (1976) 7–29. doi:10.1177/001391657600800102.
URL https://doi.org/10.1177/001391657600800102

[5] H. Leino-Kilpi, M. Välimäki, T. Dassen, M. Gasull, C. Lemonidou, A. Scott, M. Arndt, Privacy: A review of the literature, International Journal of Nursing Studies 38 (6) (2001) 663–671.

[6] P. M. Chanal, M. S. Kakkasageri, Security and privacy in iot: a survey, Wireless Personal Communications 115 (2) (2020) 1667–1693.

[7] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, Z. Lin, When machine learning meets privacy: A survey and outlook, ACM Computing Surveys (CSUR) 54 (2) (2021) 1–36.

[8] C. Zhang, Y. Tian, E. Capezuti, Privacy preserving automatic fall detection for elderly using RGBD cameras, in: International Conference on Computers for Handicapped Persons, Springer, 2012, pp. 625–633.

[9] C. Dwork, A. Roth, et al., The algorithmic foundations of differential privacy., Found. Trends Theor. Comput. Sci. 9 (3-4) (2014) 211–407.

[10] F. Pittaluga, S. J. Koppal, S. B. Kang, S. N. Sinha, Revealing scenes by inverting structure from motion reconstructions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 145–154.

[11] P. Speciale, J. L. Schonberger, S. B. Kang, S. N. Sinha, M. Pollefeys, Privacy preserving image-based localization, in: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5493–5503.

[12] P. Speciale, J. L. Schonberger, S. N. Sinha, M. Pollefeys, Privacy preserving image queries for camera localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1486–1496.

[13] K. Chelani, F. Kahl, T. Sattler, How privacy-preserving are line clouds? recovering scene details from 3d lines, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15668–15678.

[14] M. Kaur, V. Kumar, A comprehensive review on image encryption techniques, Archives of Computational Methods in Engineering 27 (2020) 15–43.

[15] R. Ayyagari, An exploratory analysis of data breaches from 2005-2011: Trends and insights, Journal of Information Privacy and Security 8 (2) (2012) 33–56.

[16] S. Subashini, V. Kavitha, A survey on security issues in service delivery models of cloud computing, Journal of network and computer applications 34 (1) (2011) 1–11.

[17] J. Byrne, B. Decann, S. Bloom, Key-nets: Optical transformation convolutional networks for privacy preserving vision sensors, arXiv preprint arXiv:2008.04469 (2020).

[18] L. S. Hill, Concerning certain linear transformation apparatus of cryptography, The American Mathematical Monthly 38 (3) (1931) 135–154.

[19] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, A. Ozcan, All-optical machine learning using diffractive deep neural networks, Science 361 (6406) (2018) 1004–1008.

[20] B. Bai, Y. Luo, T. Gan, J. Hu, Y. Li, Y. Zhao, D. Mengu, M. Jarrahi, A. Ozcan, To image, or not to image: class-specific diffractive cameras with all-optical erasure of undesired objects, eLight 2 (1) (2022) 1–20.

[21] M. Horton, S. Mehta, A. Farhadi, M. Rastegari, Bytes are all you need: Transformers operating directly on file bytes, arXiv preprint arXiv:2306.00238 (2023).

[22] Z. W. Wang, V. Vineet, F. Pittaluga, S. N. Sinha, O. Cossairt, S. Bing Kang, Privacy-preserving action recognition using coded aperture videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[23] V. Boominathan, J. T. Robinson, L. Waller, A. Veeraraghavan, Recent advances in lensless imaging, Optica 9 (1) (2022) 1.

[24] P. Latorre-Carmona, V. J. Traver, J. S. Sánchez, E. Tajahuerce, Online reconstruction-free single-pixel image classification, Image and Vision Computing 86 (2019) 28–37.

[25] K. Chelani, T. Sattler, F. Kahl, Z. Kukelova, Privacy-preserving representations are not enough: Recovering scene content from camera poses, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13132–13141.

[26] L. Ledwich, S. Williams, Reduced SIFT features for image retrieval and indoor localisation, in: Australian conference on robotics and automation, Vol. 322, Citeseer, 2004, p. 3.

[27] A. Chaari, S. Lelandais, C. Montagne, M. B. Ahmed, Global interior robot localisation by a colour content image retrieval system, EURASIP Journal on Advances in Signal Processing 2008 (2007) 1–15.

[28] E. Garcia-Fidalgo, A. Ortiz, Hierarchical place recognition for topological mapping, IEEE Transactions on Robotics 33 (5) (2017) 1061–1074.

[29] T. Sattler, Q. Zhou, M. Pollefeys, L. Leal-Taixe, Understanding the limitations of CNN-based absolute camera pose regression, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3302–3312.

[30] F. Ott, T. Feigl, C. Loffler, C. Mutschler, ViPR: visual-odometry-aided pose regression for 6DoF camera localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 42–43.

[31] H. Peter D, Kernel estimation of a distribution function, Communications in Statistics-Theory and Methods 14 (3) (1985) 605–620.

[32] B. W. Silverman, Density estimation for statistics and data analysis, Routledge, 2018.

[33] I. El Bouazzaoui, S. Rodriguez, B. Vincke, A. El Ouardi, Indoor visual slam dataset with various acquisition modalities, Data in Brief 39 (2021) 107496.

[34] N. Kuznetsov, T. Mokaev, V. Ponomarenko, E. Seleznev, N. Stankevich, L. Chua, Hidden attractors in chua circuit: mathematical theory meets physical experiments, Nonlinear Dynamics 111 (6) (2023) 5859–5887.

[35] C. Runge, Über die numerische auflösung von differentialgleichungen, Mathematische Annalen 46 (2) (1895) 167–178.