Key-Nets: Optical Transformation Convolutional Networks for Privacy Preserving Vision Sensors

Jeffrey Byrne[1]
jeff@visym.com
Brian DeCann[2]
brian.decann@stresearch.com
Scott Bloom[2]
scott.bloom@stresearch.com

[1] Visym Labs
Cambridge MA, USA

[2] Systems & Technology Research
Woburn MA, USA

Abstract

Modern cameras are not designed with computer vision or machine learning as the target application. There is a need for a new class of vision sensors that are privacy preserving by design, that do not leak private information and collect only the information necessary for a target machine learning task. In this paper, we introduce key-nets, which are convolutional networks paired with a custom vision sensor which applies an optical/analog transform such that the key-net can perform exact encrypted inference on this transformed image, but the image is not interpretable by a human or any other key-net. We provide five sufficient conditions for an optical transformation suitable for a key-net, and show that generalized stochastic matrices (e.g. scale, bias and fractional pixel shuffling) satisfy these conditions. We motivate the key-net by showing that without it there is a utility/privacy tradeoff for a network fine-tuned directly on optically transformed images for face identification and object detection. Finally, we show that a key-net is equivalent to homomorphic encryption using a Hill cipher, with an upper bound on memory and runtime that scales quadratically with a user specified privacy parameter. Therefore, the key-net is the first practical, efficient and privacy preserving vision sensor based on optical homomorphic encryption.

1 Introduction

Modern cameras are not designed with computer vision or machine learning as the target application. Security cameras are designed for professionals performing a forensic video analysis task, such that full images of a scene are collected which contain much more information about a scene than may be necessary for a target computer vision task. For example, a vision sensor with the task of face detection does not need images of nearby objects in the background, however traditional cameras collect imagery of the entire scene that can reveal much more information than intended. This is especially true for imagery collected in private spaces such as homes or businesses. Ideally, a vision sensor is privacy preserving, such that it never forms a human recoverable image that could leak information and violate end-user privacy if exposed without consent.

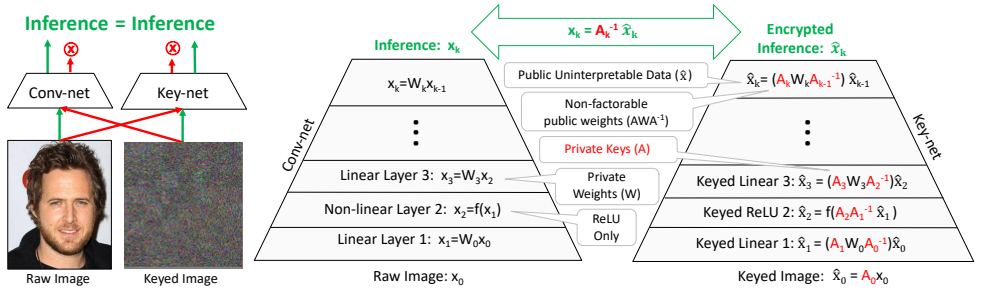

Figure 1: (left) A keynet is a convolutional network paired with a custom vision sensor which optically transforms an image such that the keynet can perform exact inference on this transformed image. The observed sensor measurement is not interpretable by a human (or any other keynet) without knowledge of the private key, which is encoded in the optics of the privacy preserving vision sensor. (right) A keynet is designed to perform inference on optical/analog encrypted inputs, by generating keyed layers from a source conv-net. The output of the keynet is equivalent to the encrypted output of the source conv-net, without ever exposing the private image of the scene or the private weights of the conv-net.

There is a need for novel design for visual sensors that are privacy preserving. Our goal is to replace traditional lens-based imaging system with a new visual sensor designed with novel diffractive or reflective optics, optimized for input to machine learning (ML) algorithms. Existing cameras have been successfully used for machine learning, however in order to protect privacy, such cameras require digital encryption and are vulnerable to exploitation through third party eavesdropping which risks making private images public. Our objective is to develop a coupled vision sensor and ML system that: (i) does not create a human interpretable image, (ii) the ML system can perform inference directly on the sensor measurements, (iii) the ML system is "keyed" so that inference can only be performed on observations from the target sensor, (iv) the parameters of the ML system are encrypted and cannot be inspected or repurposed by an adversary and (v) images are encrypted and can only be recovered with knowledge of the secret key physically encoded in the optics.

In this paper, we introduce *keynets*. A keynet is the combination of a novel vision sensor and a convolutional network designed specifically for this vision sensor. The sensor incorporates an optical transformation in the optical imaging chain that transforms an image to be uninterpretable by a human. However, the convolutional network can be designed to perform inference on this optically transformed measurement without requiring an inversion of the optical transformation. This coupling of sensor and machine learning system enforces privacy preserving vision sensing because a human interpretable image is never constructed and only the keyed convolutional network can be used for inference on this sensor. This forms a privacy preserving vision sensor that does not leak personal or private information and cannot be repurposed to another ML task.

The key contributions of this paper are:

1. *Optical transformation requirements.* We describe five sufficient conditions for an optical transformation realizable in the optics of a sensor to enable the design of a keynet from a given source convolutional network.
2. *Keynet specification.* We introduce the design of a keynet to perform inference on the

optical/analog transformation of the vision sensor, and show that a feasible family of optical transformations is based on generalized stochastic matrices. This parameterization includes a user specified privacy parameter (α).

3. *Optical element simulation*. We show that the vision sensor can be physically realized using optical 3D printed fiber bundles and analog gain preprocessing, as shown by end-to-end simulated performance.
4. *Optical homomorphic encryption.* We show that a keynet is equivalent to a homomorphic encryption based on the Hill cipher [27], which is physically realized in the optics of the vision sensor. We describe two well known weaknesses of the Hill cipher and argue that while these weaknesses are present for a generic cryptosystem, they do not introduce a practical risk for a privacy preserving vision sensor.

2 Related Work

The related work can be broadly categorized into three areas: visual sensors for machine learning, design of novel computational optics and homomorphic encryption.

First, sensors for machine learning considers redesigning vision sensors for a targeted machine learning task. An example is the movement of early layers of a convolutional network into the optical [6][16][19][59] or analog processing [12] in order to reduce power consumption. The trend towards optical processing has progressed to consider an all-optical convolutional network [39], nanophotonic CNN systems [42][70] or optical and analog CNN hybrids [15] to address the challenges of non-linearities in CNN architectures. Recently, there has also been work adapting adversarial learning principles for vision sensor design [17][37][41][44][50][51][68][69]. However, these privacy preserving approaches have demonstrated a clear (and undesirable) privacy/utility tradeoff, such that privacy increases at the expense of primary ML task performance. Our goal is to enable privacy preserving visual sensing without this privacy/utility tradeoff.

The design and fabrication of novel visual sensors considers introducing new computational optics or in-sensor analog computation suitable for a target application. For example, a coded aperture sensor [1][2][4][8][14][22][23][36] replaces the lens with phase masks realized as diffractive optical elements (DOEs) [1][4][57], such that imagery can be recovered using computational photography techniques. Such reconstruction-based approaches are also limited by a privacy/utility tradeoff [13][60]. Recent approaches attempt to eliminate the reconstruction task, but are limited by strict camera assumptions and vision tasks [20][21][67][72] or design sensors that produce partially human-interpretable images [10][31][32][47][48][49][71]. Sensor design has considered angle sensitive [16][63] or differential [62] pixels to compute precise motion or angle distribution of the light field, and single photon avalanche diodes [34][55] for ultrafast observations. Recent work on 3D printing based on two-photon lithography [64] has made possible mass production of custom optical elements at large scales. Our goal is to leverage this capability to design novel privacy preserving optical elements.

Homomorphic encryption (HE) [7][24][29][45] is a form of encryption which allows specific types of computations to be carried out on ciphertext and generate an encrypted result which, when decrypted, matches the result of operations performed on the plaintext. Homomorphic encryption has been applied to convolutional networks to perform computations on encrypted images in: CryptoNets [18][25], FHE-DiNN [9], cryptoDL [26], MiniONN [40] and Homomorphic CNNs (HCNNs) [5]. These approaches suffer from: inefficient

runtime performance, integer discretized weights [9][25], limited network depth due to increasing noise effects [5], polynomial approximations to non-linear activation layers [25] or exhibit only partially homomorphic encryption for only additive or multiplicative computations [11][29][54]. Our goal is to enable optical homomorphic encryption that does not suffer from these restrictions, but with weaker guarantees on security that we will argue is an appropriate tradeoff for a visual sensor.

3 Keynets

A *keynet* is an optically transformed convolutional network that can perform inference on data collected using a specifically designed sensor. In this section, we describe requirements for optical transformation (§3.1) and network construction (§3.2-3.4) and optical realization (§3.5). Privacy analysis is provided in supplemental material (§A.7).

Figure 1 shows an comparison of a keynet and a conv-net. In this example, there is a raw image vectorized to (x_0) which is input to a k-layer convolutional network. This network is composed as $x_k = \mathcal{N}(x_0)$, includes linear and non-linear layers, such that linear layers are represented as a sparse Toeplitz matrix (W) and the network outputs inference result x_k. The keynet uses private layer keys A_i to transform the network weights $\hat{W} = AWA^{-1}$, such that the source weights cannot be factored to recover either A or W. The keynet is paired with a custom vision sensor that physically realizes the private image key A_0 in an optical and analog transformation chain. Finally, we will show that if the non-linear layers of the source conv-net are limited to ReLU, then the keynet can operate on the transformed input $\hat{x}_k = \hat{\mathcal{N}}(\hat{x}_0)$ and $\hat{x}_k = A_k x_k$. We call this approach *optical homomorphic encryption*.

Keynets assume the following public and private information. First, the image key A_0 is secret. The physical sensor containing image key A_0 is secret, and controlled with physical security (e.g. in a locked room). The source convolutional network $\mathcal{N}$ is secret. The keyed convolutional network $\hat{\mathcal{N}}$ is public. Optically transformed images $(A_0 x)$ are public, and raw images $(A_0^{-1} A_0 x)$ are only recoverable with the secret image key. Output inference results can be either public or private (§3.4), and if private can only be recovered knowing the secret embedding key, A_k. The keyed convolutional network $\hat{\mathcal{N}}$ cannot be used to recover A or W, due to the hardness of non-negative matrix factorization (§A.7.1). Therefore, an adversary would be able to observe only the encrypted inference result $\hat{\mathcal{N}}(A_0 x)$, an uninterpretable image $(A_0 x)$ and the keyed layer weights $(\hat{W})$ but not the raw image (x_0) or the weights of the source network (W) or the raw inference (x_k).

3.1 Optical Transformation Function

Consider a family of transformation functions $\mathcal{F}$. A transformation function $f \in \mathcal{F}$ must satisfy the following five sufficient feasibility conditions to be considered an optical transformation function:

1. *Linear*. The function f must be linear $(f = A)$.
2. *Invertible*. Matrix A must be positive definite.
3. *Non-negative*. $A \geq 0$ for all matrix elements.
4. *Commutative.* There exists a non-linear activation function g that is commutative with A, such that $A(g(A^{-1}x)) = g(AA^{-1}x) = g(x)$.
5. *Sparse.* Given an $A \in \mathcal{F}$ and $B^{-1} \in \mathcal{F}$, there exists an upper bound such that for any sparse matrix W, the product AWB^{-1} is sparse with $|AWB^{-1}|_0 \leq s|W|_0$.

Condition 1 states that the transformation function must be linear, since optical image formation can be modeled as a linear transformation. Note that this is not a necessary condition, as optical propagation can include non-linear effects due to non-linear optics or diffraction. Condition 2 states that the transformation is lossless and the original image can be recovered by $A^{-1}Ax$. Condition 3 limits a matrix to be physically realizable as a linear optical element and closely connects the proposed framework with the computational complexity of non-negative matrix factorization. Condition 4 enables inference in optically encrypted convolutional networks with non-linear activation function layers. Finally, Condition 5 ensures that the end-to-end inference in the optically encrypted convolutional network is efficient and does not require the product of an infeasibly large dense matrix. A family of transformation functions $\mathcal{F}$ is defined to be an optical transformation function if all members of the family satisfy the five feasibility conditions.

3.2 Optical Transformation Convolutional Networks

Consider a convolutional network $\mathcal{N}(x)$ which is the composition of $\mathcal{N}_k$ layerwise functions:

$$\mathcal{N}(x) = \mathcal{N}_k(\mathcal{N}_{k-1}(...\mathcal{N}_1(x))) \tag{1}$$

Given an optical transformation function A and a raw image x, $\hat{x} = Ax$ is the optical transformation of the raw image into a sensor observation $\hat{x}$. First, we will consider the case where $\mathcal{N}$ is linear only, then we will extend to consider a full conv-net with non-linear layers.

Consider the case where all layers are linear. In this case, layers $\mathcal{N}_i = W$ are given by a weight matrix W which encodes the linear transformation of a trained convolutional network. For example, in a typical conv-net, linear layers include convolutional, affine, fully connected, dropout and average pooling layers. Note that since a convolution is a linear operation, it can be represented as a matrix in the form of a sparse Toeplitz matrix, where the kernel is replicated rowwise. Furthermore, multi-channel tensor inputs can be flattened to a vector x, such that the linear transformation of the layer is the matrix product of a sparse weight matrix and dense data vector. Finally, note that without loss of generality, a bias b can be applied by projective embedding $x = [x\ 1]^T$ and affine augmentation $[W\ b; 0\ 1]$. Then, the conv-net simplifies to a matrix product:

$$\mathcal{N}(x;W) = \prod_k W_k x \tag{2}$$

where the notation $\hat{N}(x;W)$ corresponds to network $\mathcal{N}$ with input x and parameters W. Given an optical transformation function A, the input Ax can be trivially input to the convolutional network as $\mathcal{N}(A^{-1}Ax)$ by inverting the data prior to inference. However, this requires exposing the image to the network. An ideal network would be able to perform inference directly on the optically transformed input, Ax, without requiring inversion.

The linear convolutional network can be constructed to operate on optical transformed inputs as follows. Linear layers can be replaced by *keyed layers* $\hat{W} = AWA^{-1}$ using secret layer keys A_i, and a secret image key A_0 such that:

$$\mathcal{N}(x;W) = A_k W_K \ldots (A_2 W_2 A_1^{-1})(A_1 W_1 A_0^{-1}) A_0 x \tag{3}$$

Recall that the keys A and inverse A^{-1} exist by conditions 1 and 2. By associativity, terms $(A^{-1}A = \mathrm{I})$ cancel, so it follows that (3) is equivalent to (2). Furthermore, by associativity,

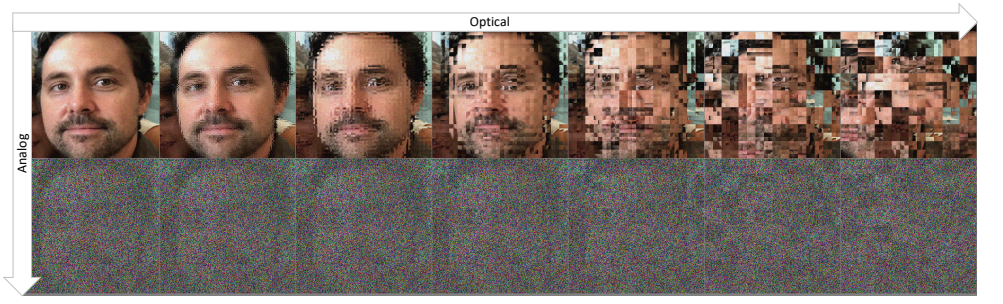

Figure 2: Generalized Doubly Stochastic Matrices combine optical and analog processing to form a human uninterpretable image (bottom right), while preserving the flexibility for partially interpretable images (middle column) that is similar to etched "optical privacy glass".

terms can be grouped into the product $\hat{W} = AWA^{-1}$. Condition 3 requires that elements of A are non-negative, which enables a proof that the factorization of $\hat{W}$ is equivalent to non-negative matrix factorization, which is NP-hard in general (§A.7.1). This protects recovery of A from $\hat{W}$. Finally, by condition 5, the product AWA^{-1} is sparse, and is at most a factor of α less sparse than W. This bounds the complexity of matrix multiplication $\hat{W}x$ to be at most α times slower than Wx. This enables a practical linear convolutional network operation that preserves sparsity of $\hat{W}$ does not require operations on an impossibly large dense matrix.

Let $\mathcal{N}(Ax;\hat{W})$ be the shorthand notation for the *keyed network* formed from keyed layers $\hat{W}$ on transformed input A_0x, such that:

$$\mathcal{N}(A_0x;AWA^{-1}) = \prod_k \hat{W}_k A_0 x \tag{4}$$

Finally, by associativity, it follows that $A_k\mathcal{N}(x;W) = \mathcal{N}(A_0x;AWA^{-1})$, since (4) is shorthand for (3) and we previously showed that (3) is equivalent to (2). This is a *homomorphism*, such that the transformed conv-net output on the original image x is equivalent to the keyed conv-net output on the optical transformed image A_0x.

Next, consider a non-linear activation function g. An activation function is a non-linear function typically used in a convolutional network structure to generate the output of a node given inputs. For example, common non-linear activation layers include rectified linear unit (ReLU), tanh and sigmoid. By condition 4, we assume that there exists a function g such that g and A are commutative:

$$Ag(A^{-1}\hat{x}) = g(AA^{-1}\hat{x}) = g(\hat{x}) \tag{5}$$

This assumption simplifies the non-linear layer to operate directly on the input, which allows the non-linear layer to be included in the keynet without modification.

Finally, any combination of linear and commutative non-linear layers can be composed into a keynet as follows.

$$\mathcal{N}_i = \begin{cases} \mathcal{N}_i(x_{i-1},\, A_i W A_{i-1}^{-1}) & \text{if } \mathcal{N}_i \text{ linear} \\ \mathcal{N}_i(x_{i-1}) & \text{if commutative non-linear} \end{cases} \tag{6}$$

This shows that the keyed network can be constructed with an optical transformation function A to enable the homomorphism $A_k\mathcal{N}(x;W) = \mathcal{N}(A_0x;AWA^{-1})$. Inference in the keyed

network on the optical transformation is equivalent to inference in the source network on the raw image. This construction enables efficient inference (condition 5) and is secure from recovery of A (condition 3). Therefore, $\mathcal{N}(Ax;AWA^{-1})$ is a homomorphic encryption scheme for inference of linearly encrypted images in a keyed conv-net. We call this *optical homomorphic encryption*.

3.3 Generalized Doubly Stochastic Matrices

Section 3.1 specified the five conditions for a feasible optical transformation function. Section 3.2 showed that a feasible optical transformation function can be used to construct a convolutional network that operates directly on optical transformed input. In this section, we show that the family of generalized doubly stochastic matrices satisfies the conditions of an optical transformation function, for choice of activation function $g = \text{ReLU}$.

A doubly stochastic matrix is defined as follows. First, a permutation matrix or *monomial matrix* Π is a square matrix that has exactly one entry of one in each row and each column and zero elsewhere, and is constructed by permuting the rows of an identity matrix. A doubly stochastic matrix is a non-negative matrix such that each row and column sums to one, that encodes a "soft" permutation. It is well known (i.e. the Birkhoff–von Neumann theorem) that every doubly stochastic matrix can be decomposed into a convex combination of permutation matrices. A generalized doubly stochastic matrix has arbitrary non-zero entries without requiring the rows and columns to sum to one. This matrix can be defined as the product of a diagonal matrix D and a doubly stochastic matrix defined as a convex combination of α permutation matrices:

$$P = \mathrm{D} \sum_{i \leq \alpha} \theta_i \Pi_i \tag{7}$$

where $\Pi = \sum_i \theta_i \Pi_i$ such that $\sum_i \theta_i = 1$, $\theta \geq 0$. The convex coefficients θ are selected to enforce that Π is positive definite. The parameter α encodes the "softness" of the stochastic matrix such that larger α is more stochastic, and $\alpha = 1$ is equivalent to a permutation matrix. Furthermore, observe that D can be extended to encode an (optional) additive bias b through an affine augmentation as $[D\ b;0\ 1]$. The term D encodes an elementwise multiplicative scaling and additive bias or *photometric* degradation, while the term Π encodes a pixelwise fractional shuffling or *geometric* degradation.

A generalized doubly stochastic matrix satisfies the five conditions of an optical transformation function (§3.1).

1. *Linear.* P is a linear function as represented by a square matrix $\mathrm{D}\Pi$.
2. *Invertible.* P is positive definite if and only if both D and Π are positive definite. A sufficient condition for P to be positive definite is selecting θ in (7) such that Π is diagonally dominant, and enforcing $\text{diag}(\mathrm{D}) > 0$.
3. *Non-negative.* Π is non-negative by construction. If D is restricted to have strictly positive diagonal entries $\text{diag}(\mathrm{D}) > 0$, then P is both non-negative and positive definite.
4. *Commutative.* Let $g(x) = \text{ReLU}(x)$ and A be restricted to a generalized permutation matrix (e.g. $\alpha = 1$ for eq. 7). Given this restriction, lemma A.1 in the supplementary material provides a proof of commutativity.
5. *Sparse.* Given an α, there exists a selection of $A \in P$ and $B^{-1} \in P$ such that the product $|AWB^{-1}|_0 \leq \alpha^2 |W|_0$, which is an upper bound on sparsity for $s = \alpha^2$. Lemma A.2 in the supplementary material provides proof of this sparsity upper bound.

3.4 Stochastic Keynets

Section 3.1 specified the five conditions for a feasible optical transformation function. Section 3.2 showed that a feasible optical transformation function can be used to construct a convolutional network that operates directly on optical transformed input. Section §3.3) showed that the family of generalized doubly stochastic matrices (e.g. "soft" permutation matrices) satisfied the conditions of an optical transformation function, for choice of activation function $g = \text{ReLU}$. A *stochastic keynet* is defined as the selection of doubly stochastic matrices for keying and ReLU for non-linear activation.

Figure 2 shows examples of generalized stochastic matrices. The horizontal scale shows optical transformations for increasingly random shuffling due to doubly stochastic matrices. The vertical scale shows analog transformations for increasingly large gains due to the diagonal matrices. The combination of these two effects results in a transformed sensor measurement in the bottom right that is uninterpretable to a human observer.

Construction of a stochastic keynet is as follows:

1. Select a pre-trained source conv-net $\mathcal{N}$ that contains only linear and ReLU layers, and a user selected privacy parameter α on $\mathcal{F}_\alpha$.
2. Randomly select a secret image key $A_0 \in \mathcal{F}_\alpha$. This is physically realized in the optical and analog imaging chain for a vision sensor as described in section 3.5.
3. If layer $\mathcal{N}_i$ is convolutional, randomly select secret layer key $A_i \in \mathcal{F}_\alpha$. Convert convolutional kernel to a sparse Toeplitz matrix and keyed layer following (3). If the convolution includes a bias term, perform an affine augmentation of the Toeplitz matrix as: $[W\ b; 0\ 1]$, with projective embedding of input tensor $[x; 1]$. If the layer includes a downsampling stride, the layer keys encode the proper shape.
4. If $\mathcal{N}_i$ is ReLU, randomly select secret layer key $A_i \in \mathcal{F}_{\alpha=1}$ such that A_i is restricted to be a scaled permutation matrix. Transform the input $g(A_i A_{i-1}^{-1} x)$.
5. If $\mathcal{N}_k$ is the output layer, select embedding key $A_k = \text{I}$ if the inference result is public data, else randomly select $A_k \in \mathcal{F}_\alpha$ if the inference is private data.
6. Compose the stochastic keynet $\hat{\mathcal{N}}(A_0 x;\ AWA^{-1})$ from $\mathcal{N}(x)$ following (6).

The stochastic keynet has two restrictions on allowable conv-net topologies. The only non-linear layer supported by this network is ReLU or ReLU-like variants (e.g. Leaky-ReLU, Parametric ReLU), as this activation function is commutative with the proposed optical transformation function. All other non-linear layers are unallowable including: max-pooling, local response normalization (LRN), sigmoid, tanh and softmax. However, all-convolutional networks have shown that replacing max-pooling with average pooling and limiting activation functions to ReLU maintains state-of-the-art performance [58]. We experimentally validate this claim in section 4.

Finally, the keynet exhibits a tradeoff between privacy and memory. A naive Toeplitz matrix construction has $O(N^2K)$ additional parameters than the source network for a layer input tensor of size (N,N) with K channels. However, these replicated channels are highly compressible. In our supplemental software, we introduce a sparse matrix format that leverages repeated submatrices as "tiles". In general, the keynet memory requirements scale as $O(\alpha^2 KT)$, given an additional tiling factor T dependent on the sparse matrix storage format. We show keynet memory requirements as a function of privacy parameter α in section 4.

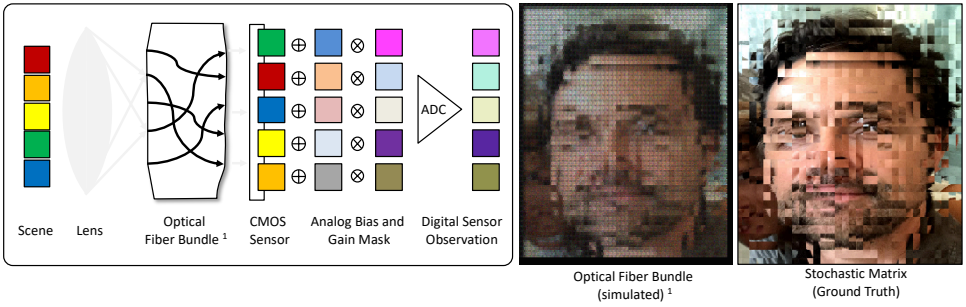

Figure 3: Simulation of a 3D printed optical fiber bundle [64] and analog preprocessing to realize a generalized doubly stochastic matrix.

3.5 Optical Realization

The sufficient conditions for an optical transform in section 3.1 define a feasible family of transformations for use in a privacy preserving vision sensor. In the supplemental material (§A.5), we show that the selected family of optical transforms based on generalized stochastic matrices can be physically realized using an analog and optical processing chain based on *3D printed incoherent fiber bundle faceplates*. An optical fiber bundle faceplate is an optical element constructed using a bundle of multi-micron-diameter optical fibers bundled into a thin plate with polished faces. A simulated example is shown in Figure 3.

4 Experimental Results

A privacy preserving vision sensor must consider the joint design of the sensor and the ML system. To justify this claim, we consider the following three experiments:

1. *Frozen System.* Does there exist an optical transformation that degrades the input image, while preserving performance of a pre-trained ML system? If such a transformation exists, then a keynet would be unnecessary, since a conv-net could be applied directly to the degraded image, and the degraded image would not be human interpretable. Experimental results show that the maximum degradation for a pre-trained network to minimize human perception [53][65] while preserving network performance is still clearly human observable. This provides evidence that preserving image privacy requires joint design of the ML system and the transformation. See supplemental material (section A.6.1), for detailed results.

2. *Trained System.* Can we jointly train an optical transformation and a conv-net to maximally degrade an image while minimizing an ML task loss? This would also render a keynet unnecessary, as fine-tuning a conv-net on degraded images would suffice. Experimental results show that jointly learning an image degradation and a network encoding using an adversarial loss can sufficiently degrade an image to render it uninterpretable by a human. However, this strategy introduces an undesirable privacy/utility tradeoff where face identification performance degrades by 12% and object detection degrades by 55%. This provides further evidence that preserving image privacy requires keying to preserve the source conv-net performance. See supplemental material (section A.6.2) for detailed results.

		Network Parameters (M)			MNIST		CIFAR-10		CIFAR-10 (sim)	
	α	LeNet	AllConvNet	VGG-16	LeNet	AllConvNet	LeNet	AllConvNet	LeNet	AllConvNet
Baseline (no privacy)	-	0.106	1.435	145.0	0.989	0.991	0.742	0.904	0.646	0.866
Keynet	2	0.111	6.131	221.7	0.989	0.991	0.742	0.904	0.646	0.866
Keynet (more privacy)	4	0.113	27.1	391.5	0.989	0.991	0.742	0.904	0.646	0.866
Keynet (most privacy)	8	0.181	100.1	1309.1	0.989	0.991	0.742	0.904	0.646	0.866

Figure 4: Keynet results. Model parameters and classification accuracy on raw and optically simulated images for a small (lenet), medium (allconvnet) and large (VGG-16) conv-net and a keynets with increasing privacy $\alpha = \{2,4,8\}$.

3. *Keyed System*. What is the simulated performance of the keynet and proposed vision sensor from Section 3.5? To demonstrate proof of concept of the proposed keynet, we have implemented key-net construction as outlined in section 3.4 in PyTorch. This prototype software exhibits exact inference performance to within floating point error. Furthermore, we simulated the keynet optical element shown in Figure 3 with simulation strategy described in detail in the supplemental material (section A.5). Results are shown in table 4 for the keynets for three baseline conv-nets.

Table 4 shows experimental for three conv-net topologies: 5-layer LeNet, an 11-layer All-Convolutional network [58] and a VGG-16 network [56]. All networks were constructed replacing max-pooling with average pooling, as per the keynet requirements. Results show the keynet memory requirements for a small (LeNet), medium (AllConvNet) and large (VGG-16) conv-net as a function of the privacy parameter (α). Naive implementation of the Toeplitz matrices in (4) results in an inefficient row-wise replication of the convolutional kernel. In our supplemental software, we introduce a new tiled sparse matrix format which provides compression of repeated submatrices. This results in a memory requirement for the keyed network on the order of 4-8x larger than the unkeyed network, depending on the selection of the privacy parameter α. Next, we trained keynets on MNIST [33] and CIFAR-10 [30], using raw images, or simulated optically transformed images from Section 3.5. Results show that the keynet achieves exact inference performance with the baseline, and that the optical simulation results in slightly degraded performance due to minor blurring of the image from fiber cross-talk. This provides proof of concept in simulation for the keynet optical element.

5 Conclusions

In this paper, we introduced keynets, which are the first practical optical homomorphic encryption scheme for the design of privacy preserving vision sensors. Our experimental results justify next steps which include: a comprehensive study of keynet semantic security as a function of privacy parameter α, GPU optimization of sparse tiled matrix-vector multiplication to improve runtime and creation and testing of a prototype optical element. Keynet software for reproducible research is available for download at https://visym.github.io/keynet. This includes two prize challenge images and public keynets for attack (§A.7.4) to encourage collaborative discovery of weaknesses in keynet security.

Acknowledgement. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001119C0067.

References

[1] Jesse K Adams, Vivek Boominathan, Benjamin W Avants, Daniel G Vercosa, Fan Ye, Richard G Baraniuk, Jacob T Robinson, and Ashok Veeraraghavan. Single-frame 3d fluorescence microscopy with ultraminiature lensless flatscope. *Science advances*, 3(12):e1701548, 2017.

[2] Nick Antipa, Grace Kuo, Reinhard Heckel, Ben Mildenhall, Emrah Bostan, Ren Ng, and Laura Waller. Diffusercam: lensless single-exposure 3d imaging. *Optica*, 5(1):1–9, 2018.

[3] Sanjeev Arora, Rong Ge, Ravi Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization - provably. In *STOC*, 2011.

[4] M. Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard Baraniuk. Flatcam: Thin, bare-sensor cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging*, 3:384 – 397, 2017.

[5] Ahmad Al Badawi, Jin Chao, Jie Lin, Chan Fook Mun, Jun Jie Sim, Benjamin Hong Meng Tan, Xiao Nan, Khin Mi Mi Aung, and Vijay Ramaseshan Chandrasekhar. The alexnet moment for homomorphic encryption: Hcnn, the first homomorphic cnn on encrypted data with gpus. In *arXiv:1811.00778v2*, 2019.

[6] Souptik Barua, Yoshitaka Miyatani, and Ashok Veeraraghavan. Direct face detection and video reconstruction from event cameras. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[7] Vishnu Naresh Boddeti. Secure face matching using fully homomorphic encryption. In *BTAS*, 2018.

[8] Vivek Boominathan, Jesse K Adams, M Salman Asif, Benjamin W Avants, Jacob T Robinson, Richard G Baraniuk, Aswin C Sankaranarayanan, and Ashok Veeraraghavan. Lensless imaging: A computational renaissance. *IEEE Signal Processing Magazine*, 33(5):23–35, 2016.

[9] Florian Bourse, Michele Minelli, Matthias Minihold, and Pascal Paillier. Fast homomorphic evaluation of deep discretized neural networks. In *IACR Cryptology ePrint Archive*, 2017.

[10] Sharmeen Browarek. *High resolution, Low cost, Privacy preserving Human motion tracking System via passive thermal sensing*. PhD thesis, Massachusetts Institute of Technology, 2010.

[11] Alon Brutzkus, Ran Gilad-Bachrach, and Oren Elisha. Low latency privacy preserving inference. In *ICML*, 2018.

[12] Mark Buckler, Suren Jayasuriya, and Adrian Sampson. Reconfiguring the imaging pipeline for computer vision. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[13] Thuong Nguyen Canh and Hajime Nagahara. Deep compressive sensing for visual privacy protection in flatcam imaging. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3978–3986. IEEE, 2019.

[14] TM Cannon and EE Fenimore. Coded aperture imaging: Many holes make light work. *Optical Engineering*, 19(3):193–283, 1980.

[15] Julie Chang, Vincent Sitzmann, Xiong Dun, Wolfgang Heidrich, and Gordon Wetzstein. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. In *Scientific Reports*, 2018.

[16] Huaijin G Chen, Suren Jayasuriya, Jiyue Yang, Judy Stephen, Sriram Sivaramakrishnan, Ashok Veeraraghavan, and Alyosha Molnar. Asp vision: Optically computing the first layer of convolutional neural networks using angle sensitive pixels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 903–912, 2016.

[17] Jiawei Chen, Janusz Konrad, and Prakash Ishwar. Vgan-based image representation learning for privacy-preserving facial expression recognition. In *arXiv:1803.07100*, 2018.

[18] Edward Chou, Josh Beal, Daniel Levy, Serena Yeung, Albert Haque, and Li Fei-Fei. Faster cryptonets: Leveraging sparsity for real-world encrypted inference. *ArXiv*, abs/1811.09953, 2018.

[19] Shane Colburn, Yi Chu, Eli Shilzerman, and Arka Majumdar. Optical frontend for a convolutional neural network. *Appl. Opt.*, 58(12):3179–3186, Apr 2019. doi: 10.1364/AO.58.003179.

[20] Ji Dai, Jonathan Wu, Behrouz Saghafi, Janusz Konrad, and Prakash Ishwar. Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 68–76, 2015.

[21] Mark A Davenport, Marco F Duarte, Michael B Wakin, Jason N Laska, Dharmpal Takhar, Kevin F Kelly, and Richard G Baraniuk. The smashed filter for compressive classification and target recognition. In *Computational Imaging V*, volume 6498, page 64980H. International Society for Optics and Photonics, 2007.

[22] R Dicke. Scatter-hole cameras for X-rays and gamma rays. *The Astrophysical Journal*, 153: L101, 1968.

[23] E Fenimore and T Cannon. Coded aperture imaging with uniformly redundant arrays. *Applied Optics*, 17(3):337–347, 1978.

[24] Craig Gentry. Fully homomorphic encryption using ideal lattices. In *STOC*, 2009.

[25] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin E. Lauter, Michael Naehrig, and John Robert Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *ICML*, 2016.

[26] Ehsan Hesamifard, Hassan Takabi, and Mehdi Ghasemi. Cryptodl: Deep neural networks over encrypted data. *ArXiv*, abs/1711.05189, 2017.

[27] Lester S. Hill. Cryptography in an algebraic alphabet. *The American Mathematical Monthly*, 36: 306–312, 1929.

[28] Kejun Huang, Nikos D. Sidiropoulos, and Ananthram Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62:211–224, 2014.

[29] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. Gazelle: A low latency framework for secure neural network inference. In *USENIX Security Symposium*, 2018.

[30] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

[31] Chiman Kwan, Bryan Chou, Jonathan Yang, Akshay Rangamani, Trac Tran, Jack Zhang, and Ralph Etienne-Cummings. Deep learning-based target tracking and classification for low quality videos using coded aperture cameras. *Sensors*, 19(17):3702, 2019.

[32] Chiman Kwan, David Gribben, and Trac Tran. Multiple human objects tracking and classification directly in compressive measurement domain for long range infrared videos. In *IEEE Ubiquitous Computing, Electronics & Mobile Communication Conference, New York City*, 2019.

[33] Yann Lecun, Leon Bottou, Y. Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278 – 2324, 12 1998. doi: 10.1109/5.726791.

[34] Changhoon Lee, Ben Johnson, and Alyosha Molnar. Angle sensitive single photon avalanche diode. *Applied Physics Letters*, 106:231105, 06 2015. doi: 10.1063/1.4922526.

[35] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[36] A. Levin, R. Fergus, F. Durand, and W. T. Freeman. Image and depth from a conventional camera with a coded aperture. In *SIGGRAPH*, 2007.

[37] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *arXiv:1904.12620v1*, 2019.

[38] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[39] Xing Lin, Yair Rivenson, Nezih T. Yardimci, Muhammed Veli, Mona Jarrahi, and Aydogan Ozcan. All-optical machine learning using diffractive deep neural networks. In *arXiv:1804.08711*, 2018.

[40] Jian Liu, Mika Juuti, Yao Lu, and N. Asokan. Oblivious neural network predictions via minionn transformations. In *ACM Conference on Computer and Communications Security*, 2017.

[41] Sicong Liu, Anshumali Shrivastava, Junzhao Du, and Lin Zhong. Better accuracy with quantified privacy: representations learned via reconstructive adversarial network. *arXiv preprint arXiv:1901.08730*, 2019.

[42] Weichen Liu, Wenyang Liu, Yichen Ye, Qian Lou, Yiyuan Xie, and Lei Jiang. Holylight: A nanophotonic accelerator for deep learning in data centers. In *Design, Automation and Test in Europe Conference and Exhibition*, pages 1483–1488, 03 2019. doi: 10.23919/DATE.2019.8715195.

[43] James R Matey, George W Quinn, Patrick Grother, Elham Tabassi, Craig Watson, and James L Wayman. Modest proposals for improving biometric recognition papers. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE, 2015.

[44] V. Mirjalili, S. Raschka, and A. Ross. Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers. *IEEE Access*, 7:99735–99745, 2019. doi: 10.1109/ACCESS.2019.2924619.

[45] Karthik Nandakumar, Nalini K. Ratha, Sharath Pankanti, and Shai Halevi. Towards deep neural network training on encrypted data. In *CVPR 2019*, 2019.

[46] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, page 6, 2015.

[47] Francesco Pittaluga and Sanjeev J Koppal. Privacy preserving optics for miniature vision sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 314–324, 2015.

[48] Francesco Pittaluga and Sanjeev Jagannatha Koppal. Pre-capture privacy for small vision sensors. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2215–2226, 2016.

[49] Francesco Pittaluga, Aleksandar Zivkovic, and Sanjeev J Koppal. Sensor-level privacy for thermal cameras. In *2016 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2016.

[50] Francesco Pittaluga, Sanjeev Koppal, and Ayan Chakrabarti. Learning privacy preserving encodings through adversarial training. In *2019 IEEE Winter Conf. Applications Comput. Vision (WACV)*, pages 791–799. IEEE, 2019.

[51] Nisarg Raval, Ashwin Machanavajjhala, and Jerry Pan. Olympus: Sensor privacy through utility aware obfuscation. In *Proceedings on Privacy Enhancing Technologies*, 2019.

[52] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *NIPS*, pages 91–99, 2015.

[53] Andras Rozsa, Manuel Günther, Ethan M. Rudd, and Terrance E. Boult. Are facial attributes adversarially robust? *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3121–3127, 2016.

[54] Theo Ryffel, Edouard Dufour Sans, Romain Gay, Francis Bach, and David Pointcheval. Partially encrypted machine learning using functional encryption. *ArXiv*, abs/1905.10214, 2019.

[55] Guy Satat, Matthew Tancik, and Ramesh Raskar. Lensless imaging with compressive ultrafast sensing. *IEEE Transactions on Computational Imaging*, 3:398–407, 2016.

[56] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[57] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (TOG)*, 37(4):114, 2018.

[58] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *ICML*, 2015.

[59] David G. Stork. Optical elements as computational devices for low-power sensing and imaging. In *Imaging and Applied Optics 2017 (3D, AIO, COSI, IS, MATH, pcAOP)*, page ITu4E.4. Optical Society of America, 2017. doi: 10.1364/ISA.2017.ITu4E.4.

[60] Christopher Thorpe, Feng Li, Zijia Li, Zhan Yu, David Saunders, and Jingyi Yu. A coprime blur scheme for data security in video surveillance. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):3066–3072, 2013.

[61] Stephen A. Vavasis. On the complexity of nonnegative matrix factorization. *ArXiv*, abs/0708.4149, 2007.

[62] A. Wang, S. Sivaramakrishnan, and A. Molnar. A 180nm cmos image sensor with on-chip optoelectronic image compression. In *Proceedings of the IEEE 2012 Custom Integrated Circuits Conference*, pages 1–4, Sep. 2012. doi: 10.1109/CICC.2012.6330604.

[63] Albert Wang, Patrick Robert Gill, and Alyosha C. Molnar. An angle-sensitive cmos imager for single-sensor 3d photography. *2011 IEEE International Solid-State Circuits Conference*, pages 412–414, 2011.

[64] Ye Wang, John Gawedzinski, Michal Pawlowski, and Tomasz Tkaczyk. 3d printed fiber optic faceplates by custom controlled fused deposition modeling. *Optics Express*, 26:15362, 06 2018. doi: 10.1364/OE.26.015362.

[65] Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli. Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13:600 – 612, 05 2004. doi: 10.1109/TIP.2003.819861.

[66] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[67] Zihao W Wang, Vibhav Vineet, Francesco Pittaluga, Sudipta N Sinha, Oliver Cossairt, and Sing Bing Kang. Privacy-preserving action recognition using coded aperture videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[68] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proc. European Conf. Comput. Vision*, pages 606–624, 2018.

[69] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Advances Neural Inf. Process. Syst.*, pages 585–596, 2017.

[70] Qiming Zhang, Haoyi Yu, Martina Barbiero, Baokai Wang, and Min Gu. Artificial neural networks enabled by nanophotonics. In *Light, science and applications*, 2019.

[71] Yupeng Zhang, Yuheng Lu, Hajime Nagahara, and Rin-ichiro Taniguchi. Anonymous camera for privacy protection. In *2014 22nd International Conference on Pattern Recognition*, pages 4170–4175. IEEE, 2014.

[72] Jinyuan Zhao. *Active Scene Illumination Methods for Privacy-preserving Indoor Occupant Localization*. PhD thesis, Boston University, 2019.