CANOPIC: PRE-DIGITAL PRIVACY-ENHANCING ENCODINGS FOR COMPUTER VISION

Jasper Tan, Salman S. Khan*[†], *Vivek Boominathan*[*], *Jeffrey Byrne*[‡], *Richard Baraniuk*[*],
Kaushik Mitra[†], *and Ashok Veeraraghavan*[*]

[*]Rice University, [†]Indian Institute of Technology Madras, [‡]Systems & Technology Research

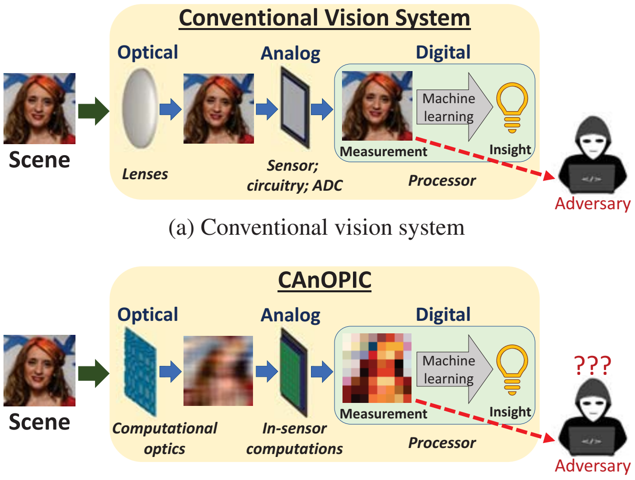
ABSTRACT

The standard pipeline for many vision tasks uses a conventional camera to capture an image that is then passed to a digital processor for information extraction. In some deployments, such as private locations, the captured digital imagery contains sensitive information exposed to digital vulnerabilities such as spyware, Trojans, etc. However, in many applications, the full imagery is unnecessary for the vision task at hand. In this paper we propose an optical and analog system that preprocesses the light from the scene before it reaches the digital imager to destroy sensitive information. We explore analog and optical encodings consisting of easily implementable operations such as convolution, pooling, and quantization. We perform a case study to evaluate how such encodings can destroy face identity information while preserving enough information for face detection. The encoding parameters are learned via an alternating optimization scheme based on adversarial learning with deep neural networks. We name our system CAnOPIC (Camera with Analog and Optical Privacy-Integrating Computations) and show that it has better performance in terms of both privacy and utility than conventional optical privacy-enhancing methods such as blurring and pixelation.

Index Terms— Privacy-preserving, computational imaging, face de-identification

1. INTRODUCTION

Recent years have seen a flurry of activity in computer vision that has resulted in a large suite of algorithms for the automatic extraction of information from the real world. Such algorithms take in an image captured by a standard camera system and extract from it the number of faces contained (*face detection*), the identity of the person in it (*face recognition*), the gestures a person is making (*gesture recognition*), and other types of beneficial information, leading to an increased integration of cameras in numerous applications. Smart security cameras are equipped with algorithms to detect when a person enters the room. Cameras in mobile phones perform face recognition to authenticate users. Video game consoles employ cameras with gesture recognition for new types of user interaction. One can expect cameras to be integrated into more devices in the coming years, devices such as wearable systems and smart clothing.



(a) Conventional vision system

(b) CAnOPIC

Fig. 1. In a conventional vision system, the optical and analog front-end passes an image of the scene to the digital sensor that is as faithful as possible, leaving the image vulnerable to digital attacks. We propose leveraging analog and optical elements to perform privacy-enhancing computations before the image is formed on the digital sensor.

However, today's cameras are digital devices: images are captured digitally and used as inputs to digital computational tools such as computer vision algorithms, leaving image data vulnerable to various digital attacks. For example, there have been numerous cases of adversaries gaining digital access to laptop cameras and baby monitors either by malware or wireless attacks [1], giving adversaries access to the digital data captured by these cameras.

Digital hacks add risks to the incorporation of computer vision for information extraction in sensitive areas, such as personal spaces and hospitals. However, for most applications, the information one wishes to extract is not in itself sensitive. The problem lies in images containing more information (including sensitive ones) than what is needed. Consider performing face detection in hospitals to track the number of people in each room. Knowing that a room contains a person may not in itself be sensitive, but knowing the person's

978-1-7281-1331-9/20/$31.00 ©2020 IEEE

identity may be. The issue is that both pieces of information are present in the standard camera image.

As an alternative, can one build a sensing device that only captures the information necessary for the desired task? Such sensors must be both private, in that they do not capture sensitive information, and useful, in that they do capture necessary information. Importantly, it is insufficient to simply perform privacy-enhancing computations digitally since these require the full digital image to begin with, allowing anyone who maliciously obtains access to the device's digital components to conceivably eavesdrop on the initial digital image.

Camera systems do naturally perform a suite of optical and analog operations before digitizing the image. Optically, lenses and other elements bend and direct incoming light rays. In the analog domain, various components operate on the voltages and currents arising from the sensor measurements. We then ask: can one leverage optical and analog elements to perform privacy-enhancing operations on the incoming light from the scene such that when the data is converted to the digital domain (for further processing), it no longer contains sensitive information but still contains information required for a vision task of interest? Such a device would be safe from all digital vulnerabilities since there is no point in the pipeline when sensitive information is encoded and stored digitally.

We refer to such systems as **CAnOPICs** (Cameras with Analog and Optical Privacy-Integrating Computations; Fig. 1). Fundamentally, this approach is different from most privacy frameworks in that it should not use any digital copy of the image. We begin the exploration of this broad field by studying a particular use case: destroying face identities in images while preserving information necessary for face detection. This is important for obtaining crowd statistics, such as foot traffic and occupancy, without compromising the identity (and thus, location) of any specific individual especially in sensitive areas such as military bases or medical settings. However, our ideas may be extended to other types of sensitive information and utility tasks. Note that in these settings, we are no longer interested in the original image but instead in the high-level information (e.g. face detection) contained in it. Multiple types of CAnOPICs can be built by using different optical technologies and analog circuitry. In this work, we propose one such CAnOPIC: 2D convolutions, max-pooling, and quantization performed in series, and we evaluate such a camera via digital simulations. While not done here, each operation can be implemented optically or in analog (discussed in Section 3 and in the Supplementary Material).

The CAnOPIC's parameters are learned via a data-driven alternating optimization method (Section 3.2). Other works have explored this method to build privacy-preserving deep neural networks (DNN), which have no straightforward implementation in the analog or optical domains. To the best of our knowledge, ours is the first work to employ this method for analog and optical operations, which pose an additional challenge as they have much smaller capacity than DNNs. Altogether, Section 4 shows that our system can greatly decrease the performance of a DNN for face identity classification while still allowing a DNN to detect whether an image contains a face or not. Just as canopic jars were used in ancient Egypt to protect their owners from the forces of nature, we propose CAnOPICs to protect sensitive data from the forces of digital attacks.

2. RELATED WORK

There exist multiple frameworks approaching privacy from different perspectives. Cryptography focuses on being able to achieve perfect reconstruction from the encrypted message, whereas we destroy sensitive information such that they indefinitely cannot be reconstructed. In this work, we explore privacy empirically, while some other works take a more theoretical approach. In [2], the authors propose a framework wherein the source distribution of a data point is kept private, which is similar to preventing classification. In [3], the authors investigate privacy of correlated data points using the Pufferfish framework, a study that may be applicable to image data due to the correlation of image pixel values.

Recently, multiple works have explored enhancing privacy via adversarial training with DNNs, a method that trains an encoder function to privatize an image against an adversarial DNN that is trained to perform the privacy-destroying task [4, 5, 6, 7, 8, 9]. These works all model the encoder as a DNN and thus require that the privacy-preserving operation be performed digitally, as there have yet to arise non-digital implementations of large-scale neural networks. We employ adversarial training to instead learn the parameters of functions that can be implemented in the analog and optical domains, giving rise to non-digital privacy-enhancers.

There are other works on imaging systems performing privacy-enhancement via optical operations. In [10], the authors design camera systems that perform blurring and k-same face de-identification in hardware. The same work also explores enhancing privacy using thermal sensors. In [11, 12], authors use coded aperture masks to enhance privacy.

3. METHOD

We present here our CAnOPIC design, consisting of a series of computations easily implementable in either the optical or analog domains. We then discuss an alternating optimization procedure to select better parameters for these computations.

3.1. CAnOPIC Design

There exist multiple optical and analog technologies that perform computations on the scene. For this work, we design our CAnOPIC to be a series of three local operations: 2D convolution, max-pooling, and quantization (illustrated in Fig. 2). Optically, 2D convolutions can be performed via diffractive

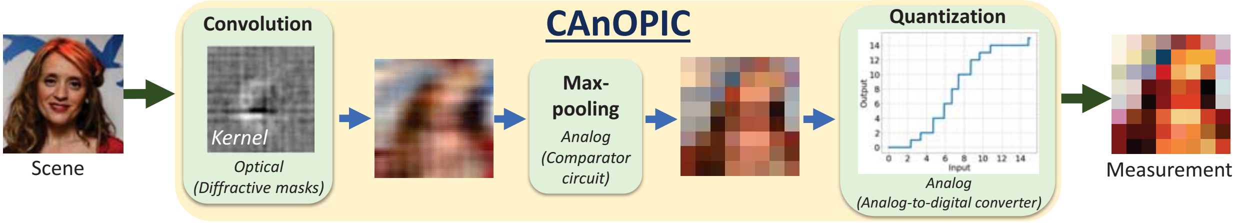


Fig. 2. We propose a combination of analog and optical components to enhance privacy in a computer vision system. It performs the following computations on an image: 2D convolution, max pooling, and quantization. The parameters of the convolution and quantization are learned in an alternating optimization procedure.

masks [13]. Max-pooling can be performed via analog comparators [14], and quantization is naturally performed by the analog-digital converter. We further discuss optical/analog implementation in the Supplementary Material.

In this work, we operate on 64 × 64 images. For the 2D convolution, we try two different schemes: (a) one 63 × 63 filter (single-filter) and (b) four 11 × 11 filters in parallel (the output being 4 RGB images). For both cases, we pad the input image by replicating the values of the edge pixels such that the image dimensions are left unchanged by the convolution. For each of the filters, the same kernel is applied to each color channel separately, which represents how diffractive masks optically perform 2D convolutions. For the max-pooling, we perform 8 × 8 max-pooling with no overlapping of patches, effectively downsizing the image by a factor of 8 in each dimension. Quantization maps real values into a finite set of possible values. While, conventionally, equally sized intervals are mapped into uniformly-spaced values, in this work, we explore intervals that need not be equally sized and output values that may not be uniformly-spaced.

3.2. Alternating Optimization to Learn Parameters

The CAnOPIC computations contain numerous parameters, such as the convolutional kernels and the quantization intervals. While these parameters can be set heuristically, one gets better performance by learning them in a data-driven fashion. Thus, we learn the convolutional kernel and quantization parameters with the following alternating optimization scheme.

The goal is to learn parameters of our CAnOPIC such that (a) face recognition cannot be performed on its outputs and (b) face detection can be performed on its outputs. We design a framework consisting of three components: (1) the CAnOPIC: 2D convolution, max-pooling, and quantization, (2) Recognition NN: a neural network (NN) trained to classify face identities, (3) Detection NN: a NN trained for the binary classification of whether an image contains a face or not. For (2) and (3), we use the 18-layer deep Residual Network (ResNet) from [15]. We use 100 identities from the VGGFace2 dataset [16] for face images and images from the ILSVRC2012 dataset [17] for the no-face class (details in Supplementary Material).

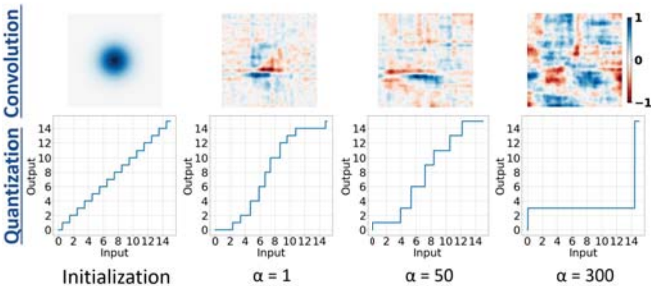


Fig. 3. Parameters learned from alternating optimization for different privacy weights α with the single-filter CAnOPIC. Notice that increasing α leads to heavier quantization.

The training is performed using two steps. In step 1, we fix the CAnOPIC and train the Recognition NN and the Detection NN to perform their classification tasks using the standard cross-entropy loss. These two networks' training are independent of each other. For step 2, we fix the Recognition NN and the Detection NN, and we then train the CAnOPIC to destroy identity information while preserving detection information using a loss function that is the sum of a privacy term and a utility term. The privacy term is set to be the negative entropy of the Recognition NN's output vector (as done in [4, 5]), which represents its predicted likelihoods/probabilities that the image belongs to each class (face identity). This term is minimized when the likelihoods are all equal, which indeed represents maximum privacy. The utility is set to the standard cross-entropy loss of the Detection NN. We also apply a multiplicative weight α on the privacy term, with higher α values emphasizing privacy over utility, to allow tweaking of the CAnOPIC's privacy-utility tradeoff. Ultimately, the encoder's loss for a clean input image x is:

$$\begin{aligned}\mathcal{L}(x,\theta_C) = \; & \text{CrossEntropy}(D(C(x)),T;\theta_C) \\ & -\alpha \text{Entropy}(R(C(x));\theta_C),\end{aligned} \quad (1)$$

where C, R and D are the CAnOPIC, Recognition NN, and Detection NN, respectively, θ_C refers to the CAnOPIC's parameters, and T refers to the target detection label.

After step 2, the encoder has learned a new set of parameters that causes the Recognition NN to fail while ensuring

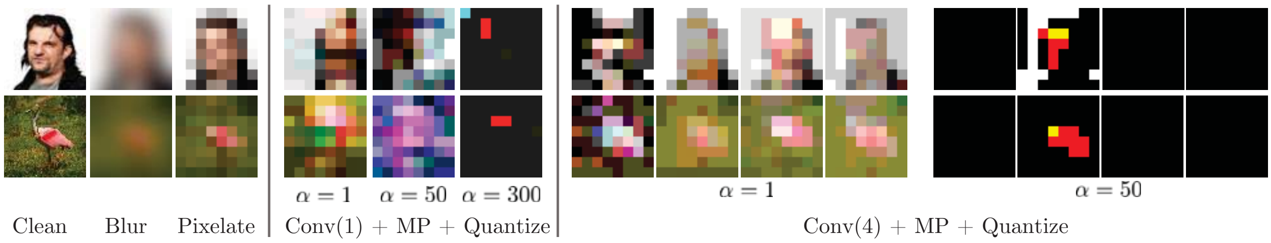


Fig. 4. Different CAnOPIC measurements for a face image (top) and a non-face image (bottom).

it does not destroy the images to the point that the Detection NN fails. However, by repeating step 1, the Recognition NN can once again learn to classify the face identities. The two steps are thus cycled through until the Recognition NN can no longer learn to classify the faces even after many training epochs. If the Recognition NN represents a strong face recognition algorithm, then this implies that the learned CAnOPIC has sufficiently destroyed face identity information.

For the single 63×63 convolutional filter case, the filter is initialized with a Gaussian kernel with a standard deviation of 8 pixels. For the case of four 11×11 convolutional filters, they are initialized with kernels where the upper half, left half, lower half, and right half are all 1's, respectively, with the remaining values being 0. Conventionally, quantization is defined by $y = \sum_{n=1}^{N-1} \mathcal{U}(x - b_i)$, where y is the discrete output, x is the continuous input, $b_i = \{0.5, 1.5, 2.5, ..., N - 1.5\}$, $N = 2^k$, k is the number of bits, and $\mathcal{U}$ is the Heaviside function. However, such formulation is not smooth and thus not suitable for backpropagation. Following [18], we approximate quantization with a differentiable version by replacing the Heaviside function with the sigmoid $\sigma()$, resulting in: $y = \sum_{n=1}^{N-1} \sigma(T(x - b_i))$, where T is a scalar hardness term gradually increased during training. The parameters learned are the b_i values. We fix the number of b_i values to be 15 and initialize them to be $0.5, 1.5, ..., 14.5$. The input to quantization is normalized to have values from 0 to 15. Learned parameters for the single-filter CAnOPIC are visualized in Fig. 3, and sample CAnOPIC measurements are shown in Fig. 4. Recall in this regime that we are not ultimately interested in the image itself but in the information we can extract from it.

4. EXPERIMENTS

In this section, we test the efficacy of our CAnOPICs in destroying face identities (privacy) while maintaining relevant information for face detection (utility) with three experiments: face identity (ID) classification, face vs. no-face classification, and detection with Faster RCNN. In these experiments, we evaluate both the single-filter and four-filter CAnOPICs (see Section 3.1) and study the effect of the privacy weight α (see Section 3.2). We also evaluate CAnOPICs both before (unlearned) and after (learned) its parameters are trained via alternating optimization, and we test different configurations of the analog/optical operations. We compare our CAnOPICs with two conventional optical privacy-enhancing methods: blurring (convolution with a Gaussian kernel with a standard deviation of 8 pixels) and pixelation (replacing each non-overlapping 8×8 patch with its mean).

Experiment 1: Face ID Classification (Privacy). We evaluate the performance of a neural network (NN) on closed-set face identity classification. The lower the performance of the NN, the higher the level of privacy of the CAnOPIC. We use the 100 identities of the VGGFace2 validation set with the largest number of images, none of which were used in the alternating optimization learning of the CAnOPIC's parameters. For each identity, 30 images are placed in the test set, 30 images are placed in a validation set, and the remaining are used for training. We pass the images through our CAnOPIC and then train the NN on the encoded measurements to minimize the standard cross-entropy loss for classification. While training our CAnOPIC via alternating optimization, we use the ResNet18 architecture for the Recognition NN. To ensure that the learned CAnOPIC is not only private for the ResNet, we use the GoogleNet architecture for this experiment.

Experiment 2: Face vs. No-Face Classification (Utility). Face detection typically consists of two components: face vs. no-face classification on different patches and refinement of bounding box coordinates. We show CAnOPIC measurements still contain sufficient information for face detection by focusing on the first component. We train a GoogleNet to classify whether an image contains a face or not. For training data, we use 100,000 images from VGGFace2 and 100,000 faceless images from the ILSVRC2012 training set for the face and no-face classes, respectively. For testing, we use the same test set from the ID classification experiment for the face class and 3,000 faceless images from the ILSVRC2012 validation set for the no-face class.

Results for experiments 1 and 2 are shown in Table 1 and Fig. 5, which reveal some noteworthy insights. First, the result for the 4-filter CAnOPIC with $\alpha = 1$ shows that we have designed a system universally better than the standard blurring and pixelation methods: it is both more private (lower ID classification score) and more useful (higher face vs. no-face score). Second, both CAnOPICs for the $\alpha = 1$ case show the success of our alternating optimization procedure in training the CAnOPICs' parameters to yield both a more private and more useful system. Third, the results for the different privacy

Table 1. Results for the ID classification and face vs. no-face classification tasks for different configurations. Conv(1) refers to a single 63×63 convolutional filter while Conv(4) refers to four 11×11 convolutional filters. MP refers to max-pooling.

	Unlearned		**Learned**	
	ID Classification	**Face vs. No-Face**	**ID Classification**	**Face vs. No-Face**
Clean (no encoding)	97.60%	99.81%	X	X
Gaussian Blur	73.50%	96.61%	X	X
Pixelate (average pool)	75.73%	97.73%	X	X
Conv(1) + MP + Quantize ($\alpha = 1$)	36.70%	93.51%	**33.73%**	**94.66%**
Conv(1) + MP + Quantize ($\alpha = 50$)	36.70%	93.51%	19.13%	91.87%
Conv(1) + MP + Quantize ($\alpha = 300$)	36.70%	93.51%	5.2%	74.81%
Conv(4) + MP + Quantize ($\alpha = 1$)	72.13%	96.94%	**66.50%**	**98.44%**
Conv(4) + MP + Quantize ($\alpha = 50$)	72.13%	96.94%	17.87%	87.75%
Conv(1) ($\alpha = 300$)	73.50%	96.61%	56.17%	97.03%
MP	60.40%	96.88%	X	X
Quantize ($\alpha = 300$)	97.13%	99.81%	8.63%	78.03%
Conv(1) + MP ($\alpha = 300$)	54.40%	94.49%	24.17%	92.50%
Conv(1) + Quantize ($\alpha = 300$)	59.43%	95.36%	12.07%	84.47%
MP + Quantize ($\alpha = 300$)	51.57%	96.01%	6.73%	74.56%

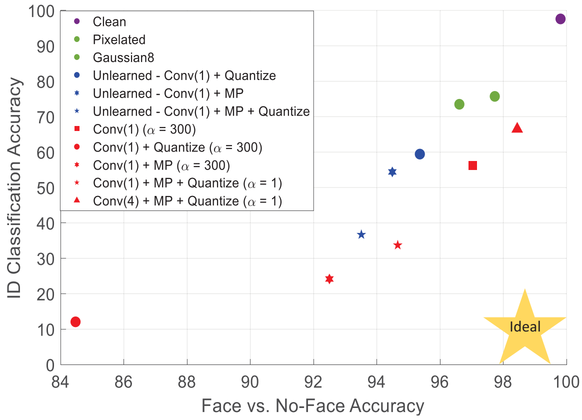
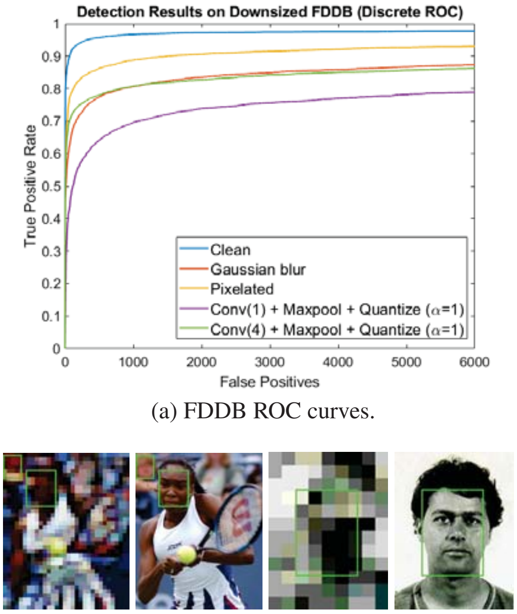


Fig. 5. ID classification and face vs. no-face results. Green: conventional methods; blue: unlearned CAnOPICs; red: learned CAnOPICs. Learned CAnOPICs bring us closer to the ideal privacy-enhancing encoder.



(a) FDDB ROC curves.

(b) Sample encoder predicted bounding boxes.

Fig. 6. (a) FDDB discrete ROC results [19] for clean, blurred, pixelated, and CAnOPIC images. (b) Predicted bounding boxes for the Conv(1) + Max-pool + Quantize ($\alpha = 1$) CAnOPIC overlaid on both the encoded and original images.

weights α show how one can tweak α to obtain CAnOPICs at different points of the privacy-utiltiy tradeoff, allowing one to choose from a continuum of possible encoders depending on the privacy requirements of the application.

Experiment 3: Detection with Faster RCNN (Utility). We also test CAnOPICs for the full task of face detection (i.e. providing bounding box coordinates) using a standard Faster R-CNN detection network [20] trained on the WIDER face dataset [21] passed through the CAnOPIC and tested on the FDDB face detection benchmark [19] passed through the CAnOPIC. The CAnOPIC's privacy-enhancing effect depends on the size of the face since the convolutional and max-pooling kernels have fixed sizes. In the ID classification experiments, the size of a face is roughly 40 pixels in its larger dimension. Larger faces would be less private since the CAnOPIC's effect is weaker on them. For fairness, we downsample each FDDB image such that its smallest face has a major axis diameter of at most 40 pixels (FDDB face annotations are ellipses). Images with faces smaller than 40 pixels are left at the same size. We test our two $\alpha = 1$ CAnOPICs (single-filter and 4-filters), blurring, and pixelation and present the results in Fig. 6. The CAnOPICs detect a reasonably large number of faces, but their performances are slightly poorer than those of blurring and pixelation. This may be because the CAnOPICs were not optimized for the full task of face detection, and we may not be accounting for functional differences between face vs. no-face classification (wherein our CAnOPIC performs better) and detection. Refining our method to design CAnOPICs more suited for full face detection is a potential future direction of this work.

5. CONCLUSIONS

In this work, we introduce the problem of preserving privacy on images by performing operations before any digital conversions. To this end, we propose CAnOPICs: cameras with analog and optical privacy-integrated computations. We propose one specific system of performing 2D convolution, max pooling, and quantization in series, each of which can be easily implemented in either the optical or analog domains. We also present an alternating optimization scheme to learn the various parameters of the CAnOPIC. Our results show that our system causes neural networks to fail in performing tasks related to identifying the human faces (even if they can train on CAnOPIC measurements) while still allowing face detection to be performed with reasonable accuracy. It is our hope that our framework of destroying sensitive information takes us one step closer towards applicable privacy-preserving computer vision sensors safe from digital vulnerabilities.

Acknowledgement. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001119C0067 and by NSF grants CCF-1911094, IIS-1838177, IIS-1652633, and IIS-1730574; ONR grants N00014-18-1-2047, N00014-18-12571 and N00014-17-1-2551; AFOSR grant FA9550-18-1-0478; DARPA grant G001534-7500; and a Vannevar Bush Faculty Fellowship.

6. REFERENCES

[1] R. Borison, “97 people arrested for hacking into webcams remotely and spying on people,” May 2014.

[2] J. Geumlek and K. Chaudhuri, “Profile-based privacy for locally private computations,” *arXiv preprint arXiv:1903.09084*, 2019.

[3] S. Song, Y. Wang, and K. Chaudhuri, “Pufferfish privacy mechanisms for correlated data,” in *Proc. 2017 ACM Int. Conf. Manag. Data*. ACM, 2017, pp. 1291–1306.

[4] Z. Wu, Z. Wang, Z. Wang, and H. Jin, “Towards privacy-preserving visual recognition via adversarial training: A pilot study,” in *Proc. European Conf. Comput. Vision*, 2018, pp. 606–624.

[5] P. C. Roy and V. N. Boddeti, “Mitigating information leakage in image representations: A maximum entropy approach,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2019, pp. 2586–2594.

[6] C. Huang, P. Kairouz, and L. Sankar, “Generative adversarial privacy: A data-driven approach to information-theoretic privacy,” in *IEEE Asilomar Conf. Signals, Syst., Comput.*, 2018, pp. 2162–2166.

[7] F. Pittaluga, S. Koppal, and A. Chakrabarti, “Learning privacy preserving encodings through adversarial training,” in *2019 IEEE Winter Conf. Applications Comput. Vision (WACV)*. IEEE, 2019, pp. 791–799.

[8] S. Liu, A. Shrivastava, J. Du, and L. Zhong, “Better accuracy with quantified privacy: representations learned via reconstructive adversarial network,” *arXiv preprint arXiv:1901.08730*, 2019.

[9] V. Mirjalili, S. Raschka, and A. Ross, “Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers,” *IEEE Access*, vol. 7, pp. 99735–99745, 2019.

[10] F. Pittaluga and S. J. Koppal, “Pre-capture privacy for small vision sensors,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2215–2226, 2017.

[11] Z. W. Wang, V. Vineet, F. Pittaluga, S. N. Sinha, O. Cossairt, and S. B. Kang, “Privacy-preserving action recognition using coded aperture videos,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognition Workshops*, 2019.

[12] T. N. Canh and H. Nagahara, “Deep compressive sensing for visual privacy protection in flatcam imaging,” in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, 2019.

[13] V. Boominathan, J. Adams, J. Robinson, and A. Veeraraghavan, “Phlatcam: Designed phase-mask based thin lensless camera,” *IEEE Trans. Pattern Anal. Mach. Intell.*, in-press.

[14] E. G. Nestler, M. M. Osqui, and J. G. Bernstein, “Convolutional neural network,” June 2017, US Patent App. 15/379,114.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conf. Comput. Vision Pattern Recognition*, June 2016, pp. 770–778.

[16] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “Vggface2: A dataset for recognising faces across pose and age,” in *Int. Conf. Automat. Face Gesture Recognition*, 2018.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., “Imagenet large scale visual recognition challenge,” *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[18] J. Yang, X. Shen, J. Xing, X. Tian, H. Li, B. Deng, J. Huang, and X. Hua, “Quantization networks,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2019, pp. 7308–7316.

[19] V. Jain and E. Learned-Miller, “Fddb: A benchmark for face detection in unconstrained settings,” Tech. Rep. UM-CS-2010-009, Univ. Massachusetts, Amherst, 2010.

[20] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[21] S. Yang, P. Luo, C. C. Loy, and X. Tang, “Wider face: A face detection benchmark,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2016.