Research Article | Vol. 33, No. 12 / 16 Jun 2025 / Optics Express 26590

Optics EXPRESS

Co-designed pre-capture privacy optics for computer vision

CARLOS MAURICIO VILLEGAS BURGOS,[1] YU FENG,[2] PEI XIONG,[1] YUHAO ZHU,[2,3] AND A. NICKOLAS VAMIVAKAS[1,4]

[1] University of Rochester, Institute of Optics, 275 Hutchison Road, Rochester, NY 14627, USA
[2] University of Rochester, Department of Computer Science, 2513 Wegmans Hall, Rochester, NY 14627, USA
[3] yzhu@rochester.edu
[4] nick.vamivakas@rochester.edu

Abstract: A metasurface-based pre-capture privacy optical system is jointly optimized with a binary classification computer vision (CV) task. The pre-capture privacy module degrades the quality of images before they are captured and saved to the device's memory, making these obscured inputs the only available data for both privacy-breaching attackers and the CV task model. The objective of attaining a trade-off of preserving privacy and CV performance that favors the latter was accomplished as a demonstration of the joint optimization co-design scheme proposed in this work. Specifically, when comparing the maximal performance of the attacker and CV task models attained on their respective tasks when taking unobscured versus obscured inputs, the former model's performance drastically drops from 75.4% accuracy to 17.5% accuracy, while the latter's decreases from 95.7% average precision (AP) to 67.0% AP.

1. Introduction

Recent decades have seen widespread application of computer vision technologies over a multitude of fields. As a consequence of the ensuing ubiquitous deployment of digital cameras that continuously collect image data, concerns about privacy have been growing, and there has been an increasing demand to address them [1]. While many different approaches exist to conceal sensitive information contained in the captured images, in this work we will solely focus on the one where images are obscured or encrypted before they are captured by the camera's sensor, referred to as "pre-capture privacy" [2–5]. The merit of this approach is to prevent malicious attackers from ever accessing the unobscured version of the captured images, since only the obfuscated version is saved in digital form to the device's memory. However, the downside is that these obfuscated images are also the only ones that are available for the device's computer vision task to work with, which can lead to a decrease in its performance. In order to mitigate this, some approaches have been proposed to use deep learning algorithms [6,7] to co-train (jointly optimize) the parameters of the pre-capture privacy module with those of the computer vision task's computational model [4,5]. In these, an adversarial computational model that attempts to extract the sensitive information from the obfuscated images is integrated into the co-training scheme, with the purpose of making the system produce obfuscations that are more robust against sophisticated privacy-breaching attacks that make use of deep learning algorithms. The goal of these joint optimization techniques is to reach a favorable trade-off between privacy preservation and the computer vision task's performance.

The emerging field known as "Deep Optics" is characterized by the usage of deep learning algorithms to optimize the parameters of optical systems designed to perform domain-specific tasks [8–10]. Previous works in this field have demonstrated various optimized optical systems designed to carry out a wide variety of computer vision task applications [11–15]. Furthermore, there are existing works that demonstrate the design of different types of optical elements that were optimized to introduce pre-capture privacy while maintaining the performance of a given

#557638 https://doi.org/10.1364/OE.557638
Journal © 2025 Received 23 Jan 2025; revised 15 Apr 2025; accepted 4 May 2025; published 13 Jun 2025

Optics EXPRESS

computer vision task. Specifically, these works use coded apertures with pinhole arrays used in lensless cameras [16,17], or a phase mask on an imaging system with lenses [18–20]. A common approach present in all of them is the usage of a single customized optical element that tailors the point spread function (PSF) of the device's imaging system to degrade the entirety of the captured images. Alternatively, some recent works have proposed a different approach for designing optical systems that implement pre-capture privacy, where a set of optimized diffractive layers are used to attain more complex functionality, such as optically suppressing some classes of objects in the captured scene while others are imaged with high quality [21,22]. While this approach opens up avenues for more sophisticated optical encryption than the preceding works, it has currently only been demonstrated in the terahertz (THz) range and in experiments where only images of simple objects (monochromatic, low-resolution handwritten digits) are being captured.

In this work, we demonstrate a co-trained imaging system that works at visible wavelengths and that attains a favorable trade-off between privacy preservation and a good level of performance on an object classification computer vision task. Furthermore, the input images are taken from a more complex dataset than the one used in our previous work [23], and they consist of high-resolution pictures of different classes of objects [24]. These images are degraded by the aberrations introduced into the PSF of the imaging system before they are captured by the camera's sensor. The PSF is controlled by a phase modulation profile that a co-designed metasurface imparts on incident light. Compared to the previous pre-capture privacy works that also use phase masks [18–20], this work attains the physical implementation of a phase profile composed by a much larger number of Zernike polynomials. Specifically, the phase-modulating optical element that was used in our laboratory experiments realized a phase profile with 252 Zernike polynomials, while the previous works were limited to only 15 Zernike polynomials. The increased parameter space allows the optical system in this work to produce more severe optical aberrations that further degrade the quality of the captured images.

A metasurface is our phase-modulating optical element of choice because of the flexible manipulation of light's properties with subwavelength resolution that it offers, along with its lightweight and compact form factor [25,26]. Specifically, the phase modulation profile is implemented via the geometrical phase that is introduced by the light's interaction with an array of anisotropic nano-pillars [25,27]. By using unit cells containing multiple nano-pillars that each have different geometrical properties, it is possible to create independent phase modulation profiles for different wavelength bands in the visible spectrum [28,29]. Conversely, it is also possible to multiplex different phase modulation profiles (each with a distinct constant amplitude modulation profile) within the metasurface's area. This work focuses on exploring the latter possibility, using the PSF yielded by three multiplexed phase modulation profiles to obfuscate monochromatic input images in both laboratory experiments and simulations. Complementarily, the former possibility is still studied with secondary simulations, similarly to our previous work [29], to apply independent PSFs on each color channel of input RGB images. Finally, this work is the first experimental demonstration of a geometrical phase metasurface being jointly designed with a privacy-aware computer vision algorithm for image classification.

2. System design

2.1. Joint optimization scheme

We use the three-way co-training scheme illustrated in Fig. 1(a), which involves the following parametrized, differentiable computational models: The optical system's image formation model (referred to as "Optics model" for simplicity), the computer vision (CV) task model, and an adversarial attacker model (referred to as simply "Attacker model" for the rest of this work). Under this scheme, each model is trained (optimized) via gradient descent using deep learning algorithms. This means that performance-related loss functions are used to iteratively update

Optics EXPRESS

the models' parameters after computing the gradient of these loss functions with respect to said parameters.

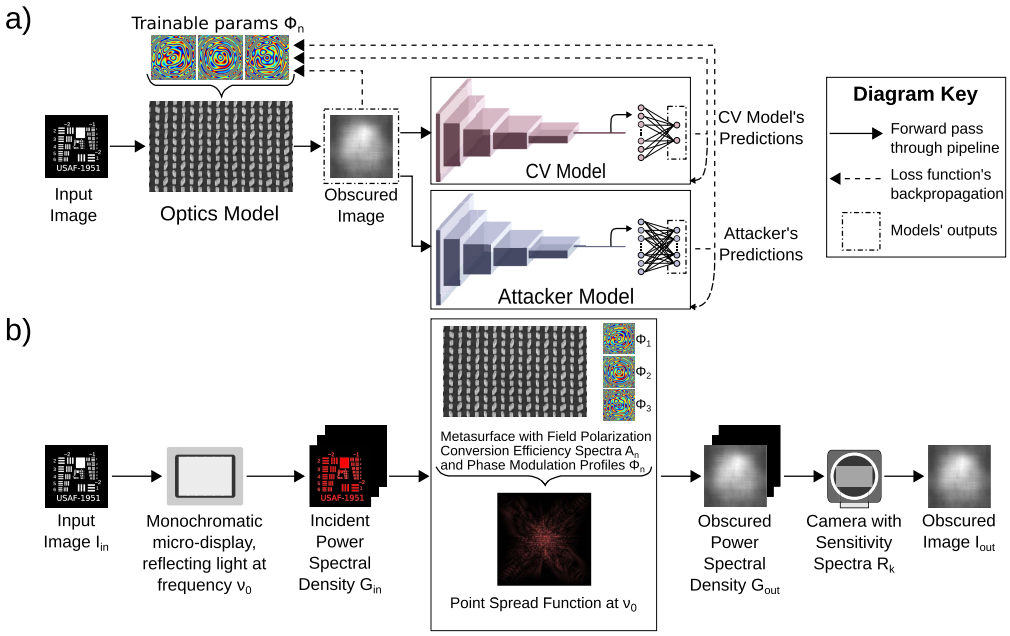

Fig. 1. a) Co-training scheme: A differentiable, parametrized, computational Optics model that introduces the optical aberrations is jointly optimized with a Computer Vision (CV) and an Adversarial (Attacker) deep learning models. By design, the Optics model's optimization process receives feedback from those of the other two models, with the objective of stopping the Attacker's attempts to extract sensitive private information from the obscured input images, without hindering the performance of the CV model's task. b) Schematic of the Optics model's forward pass. The model estimates the wavelength channels in the input scene's power spectral density, and simulates the interaction of these components with the metasurface and the camera's sensor as light of different wavelengths propagates through the system. Since a monochromatic display is used to project images in this work's laboratory experiments, only one wavelength component is of interest, as illustrated in this diagram.

The Optics model has the goal of degrading the quality of the input images before they are passed down as inputs for the CV task and the Attacker models. The loss function L_{Opt} quantifies the extent to which this goal is accomplished. Meanwhile, the CV task and Attacker models are optimized independently of each other using loss functions L_{CV} and L_{Atk}, respectively, with the goal of maintaining their performance in spite of the increasing amount of optical aberrations present in the images they receive as inputs. A more in-depth explanation about the design of these loss functions can be found in the Supplement 1 document.

The loss functions L_{CV} and L_{Atk} are included as terms in the functional form of L_{Opt}, which translates into the Optics model receiving feedback from the other two models, and which results into a coupling between the training processes of the three models. By design, when updating the Optics model's parameters by minimizing L_{Opt} via gradient descent, the L_{Atk} term is driven to increase while the L_{CV} term is driven to decrease. This makes the Optics model converge to a state where it produces optical aberrations that drastically reduce the performance of the Attacker model without significantly hindering the CV task model. This way, a favorable trade-off between the preservation of both pre-capture privacy and computer vision task performance can be attained.

Optics EXPRESS

2.2. Image formation model

The Optics model is built to simulate the image formation process of the set-up that would later be used in the laboratory experiments. An overview of the computations performed to run the Optics model is presented by the diagram found in Fig. 1(b). The optical system layout that implements this pipeline consists of a 4*f* system [30] where a micro-display is placed in its input plane, a geometrical phase metasurface is placed in its Fourier plane, and an output image is formed in its output plane. Additionally, an arrange of polarization-manipulating optical elements are used to make the light incident on the metasurface have a circular polarization state. A diagram of the laboratory experiment's set-up can be found in Fig. 2. Furthermore, a more detailed explanation and the specifications of this system's components are found in Section 3.3.

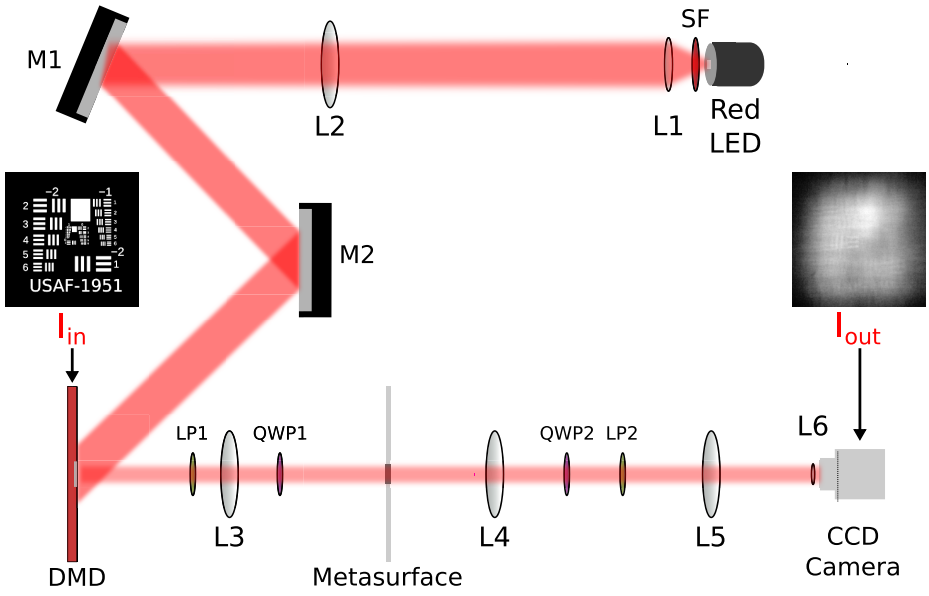

Fig. 2. Light emitted by a red LED is spectrally filtered by a band-pass filter SF, is then expanded and collimated by lenses L1 and L2, and is finally steered towards the digital micro-mirror display (DMD) by mirrors M1 and M2. Images projected by the illuminated DMD go into the 4*f* system comprised by lenses L3 and L4, in whose Fourier plane the fabricated metasurface is placed. Light is converted into the left-hand circular polarization state by linear polarizer LP1 and quarter-wave plate QWP1 before being incident on the metasurface. Their counterparts QWP2 and LP2 filter out any unconverted light, leaving only the right-hand circularly polarized component encoding the obscured image, which is shrunk down and relayed to the CCD camera's sensor by the 4*f* system comprised by lenses L5 and L6.

When circularly polarized light is transmitted by an anisotropic element that has an in-plane orientation angle θ, a wavelength-dependent portion of the light is converted to the circular polarization state of opposite handedness and is imbued with a wavelength-independent geometrical phase equal to 2θ [27]. The metasurface is comprised by an array of anisotropic nano-pillars with rectangular cross-sections. The polarization conversion efficiency (PCE), which is the fraction of the incident light that gets imbued with the geometrical phase, is determined by the geometrical properties of the nano-pillar that is interacting with the incident light [28]. By varying the geometrical parameters and in-plane orientation of the nano-pillars located at each coordinate (u, v) in the metasurface's plane, we can produce a customizable amplitude and phase

Optics EXPRESS

modulation profile represented by the frequency-dependent transmission function $T(u,\, v;\, \nu)$, where ν denotes the frequency dependence.

It can be shown that for a linear imaging system with spatially incoherent illumination, the power spectral density $\mathcal{G}_{\text{out}}(x,\, y;\, \nu)$ in the output plane is given by a convolution between the power spectral density $\mathcal{G}_{\text{in}}(\xi,\, \eta;\, \nu)$ in the input plane and a frequency-dependent point spread function $\text{PSF}(\xi,\, \eta;\, \nu)$:

$$\mathcal{G}_{\text{out}}(x,\, y;\, \nu) = \mathcal{G}_{\text{in}}(\tilde{\xi},\, \tilde{\eta};\, \nu) * \text{PSF}\left(\tilde{\xi},\, \tilde{\eta};\, \nu\right) = \iint \mathcal{G}_{\text{in}}(\tilde{\xi},\, \tilde{\eta};\, \nu)\text{PSF}(x-\tilde{\xi},\, y-\tilde{\eta};\, \nu)d\tilde{\xi}d\tilde{\eta}. \quad (1)$$

Furthermore, in the case of a $4f$ system, the frequency-dependent PSF is given by the magnitude squared of the Fourier transform of the transmission function $T(u,\, v;\, \nu)$ of the optical element placed in its Fourier plane:

$$\text{PSF}(x-\xi,\, y-\eta;\, \nu) = \left|\mathcal{F}_{\text{2D}}\left\{T(u,\, v;\, \nu)\right\}\right|^2\Big|_{\left(\frac{\nu}{c_0/2}(x-\xi),\, \frac{\nu}{c_0/2}(y-\eta)\right)}, \quad (2)$$

where c_0 is the speed of light in vacuum and f_2 is the focal length of the second lens in the $4f$ system. The Supplement 1 document contains a derivation of the above equations, where the optically anisotropic properties of the metasurface placed in the Fourier plane of the $4f$ system used in this work are taken into account.

Additionally, we model the power spectral density produced by the micro-display in terms of the emission spectra $S_c(\nu)$ of its pixels, where $c = \{\text{R}, \text{G}, \text{B}\}$ denotes the c-th color channel of the projected RGB images. The power spectral density produced by the display to project the images' c-th color channel can be represented as a separable function, i.e. as a product between the function that only depends on frequency, $S_c(\nu)$, and a function that only depends on spatial position, $I_{\text{in},c}(\xi,\, \eta)$. Assuming that the light emitted by the display is incoherent, i.e. each pixel is statistically independent from the others, we can model the total power spectral density produced by the display to project an input image as the sum of the power spectral densities associated with each color channel of said image:

$$\mathcal{G}_{\text{in}}(\xi,\, \eta;\, \nu) = \sum_c S_c(\nu) I_{\text{in},c}(\xi,\, \eta), \quad (3)$$

where $I_{\text{in},c}(\xi,\, \eta)$ represents the c-th color channel of the input digital RGB image that is being projected by the display.

Finally, the raw signal intensity from the k-th type of camera pixel (the k-th channel of the raw RGB image) in the detector's plane (the system's output plane) is given by the integral over frequencies of the product between the incident power spectral density $\mathcal{G}_{\text{out}}(x,\, y;\, \nu)$ and the spectral sensitivity $R_k(\nu)$ of the pixel located at point $(x,\, y)$:

$$I'_{\text{out},k}(x,\, y) = \int_0^\infty R_k(\nu)\mathcal{G}_{\text{out}}(x,\, y;\, \nu)d\nu\,. \quad (4)$$

In this work's main experiments, we used a monochromatic illumination source and a micro-display with reflective pixels to project the input images. As such, we model the input power spectral density from Eq. (3) as:

$$\mathcal{G}_{\text{in}}(\xi,\, \eta;\, \nu) = \delta(\nu-\nu_0) \sum_c \tfrac{1}{3} I_{\text{in},c}(\xi,\, \eta), \quad (5)$$

where ν_0 is the frequency of the monochromatic light that illuminates the reflective micro-display. In other words, the $S_c(\nu)$ from Eq. (5) were modeled as delta functions centered on ν_0 for the

Optics EXPRESS

three color channels of the input RGB images. The way in which Eq. (5) is written shows that this is equivalent to projecting a grayscale monochromatic image obtained by taking the average between color channels from every pixel of the original RGB digital inputs.

In a similar fashion, the final digital image $I_{\text{out}}(x, y)$ that serves as the output of the Optics model and is passed down as an input to the CV task and Attacker models is obtained from $I'_{\text{out}}(x, y)$ by taking the average over the color channels of the image processed by the camera's image signal processing (ISP) algorithm. This is done because both the CV task and Attacker models expect to take grayscale monochromatic images as inputs by design. More details about how a simulation of the camera's ISP algorithm was incorporated into our Optics model can be found in the Supplement 1 document.

Additionally, we conducted complementary simulations where color input images were projected with an OLED display. These made use of Eq. (3) instead of Eq. (5), and the process for obtaining $I_{\text{out}}(x, y)$ from $I'_{\text{out}}(x, y)$ still made use of the simulated camera's ISP algorithm. However, the final color-averaging step before passing $I_{\text{out}}(x, y)$ to the CV task and Attacker models was omitted, since those models expected color images as inputs in that case.

2.3. Optics model parametrization

The image formation model makes use of the following fixed parameters: The spectra of the light coming from the micro-display's pixels $S_c(\nu)$, the sensitivity spectra of the camera's pixels $R_k(\nu)$, and the field PCE spectra $A_n(\nu)$ of the metasurface's nano-pillars, all of which are shown in Fig. 3(a). As mentioned in Section 2.2, $S_c(\nu)$ are modeled as delta functions centered on a design frequency ν_0 in the main experiments of this work. More specifically, the value of ν_0 is $\nu_0 = c_0/(632.8\,\text{nm}) \approx 474\,\text{THz}$. The emission spectra $S_c(\nu)$ of the OLED micro-display's pixels (used in the complementary simulations of the broadband illumination case) and the camera's sensitivity spectra $R_k(\nu)$ were obtained from technical specifications provided by the devices' manufacturers. Meanwhile, the metasurface's field PCE spectra $A_n(\nu)$ were obtained by finite-difference time-domain (FDTD) simulations of the interactions between its nano-pillars and incident light. Further information about these FDTD simulations can be found in Section 3.2. It should be noted that it is possible to have the properties of the nano-pillars (along with the yielded PCE spectra) be optimizable parameters rather than fixed ones [31–34], but taking that approach is out of the scope of this work.

The trainable (optimizable) parameters of the image formation model are the phase modulation profiles $\Phi_n(u, v)$ that introduce optical aberrations into the imaging system, which are produced by the arrays of nano-pillars in the metasurface. These phase modulation profiles are two-dimensional numerical arrays that are parametrized as a linear combination of Zernike polynomials:

$$\Phi_n(u, v) = \sum_{j=1}^{J} \alpha_{n,j} Z_j(u, v), \tag{6}$$

where $Z_j(u, v)$ is the j-th Zernike polynomial in OSA notation [35,36], and $\alpha_{n,j}$ is the j-th scalar coefficient that is used to build the n-th phase modulation profile. In this work, we make use of $J = 252$ terms in the Zernike polynomial basis to parametrize $N = 3$ distinct phase modulation profiles. Each of these three phase modulation profiles is imparted on incident light by an array of nano-pillars that have shared geometrical properties but different in-plane orientation angles. The metasurface layout is divided into unit cells that contain these three types of nano-pillars, so the three modulation profiles are spatially multiplexed. The layout and dimensions of the nano-pillars placed in these unit cells is shown in Fig. 3(b).

The size of the metasurface's unit cells is in the order of one wavelength, while the dimensions of the nano-pillars and the separation between them are sub-wavelength. This fact is taken into account when implementing the numerical arrays that represent the metasurface's modulation

Optics EXPRESS

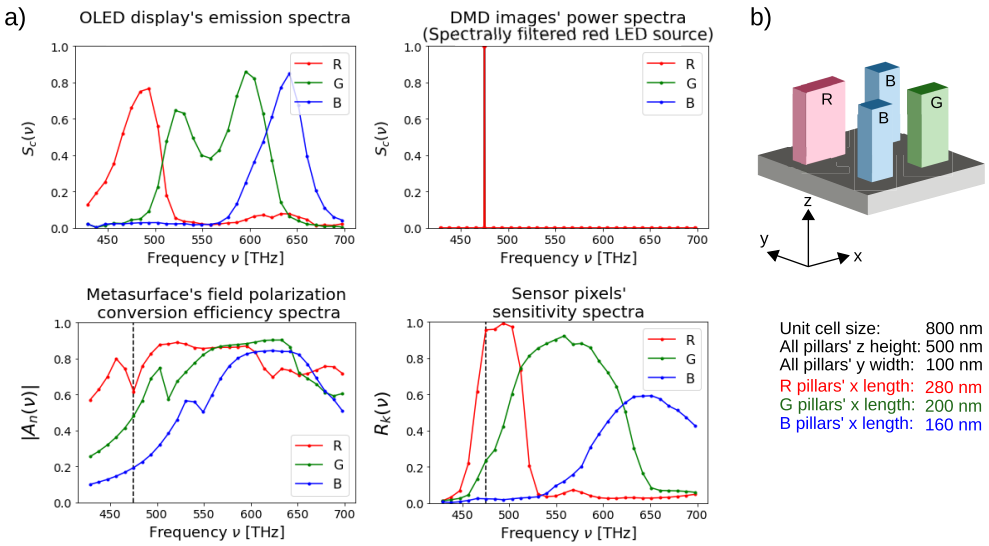

Fig. 3. a) Plots of the data values of the Optics model's fixed parameters: the light sources' emission spectra S_c, the metasurface nano-pillars' field polarization conversion efficiency spectra A_n, and the sensor pixels' sensitivity spectra R_k. b) Layout and geometrical specifications of the metasurface's unit cells, which contain three types of nano-pillars. The RGB-based color coding of the nano-pillars in this diagram matches the label of their corresponding A_n curve plotted in a).

profiles, which is necessary to properly compute the Fourier transform operations. The metasurface's transmission is modeled as a continuous function given by the sum of the spatially-multiplexed transmission profiles associated with the arrays of the three different types of nano-pillars:

$$T(u,\, v;\, \nu) = \sum_{n}^{N} A_n(\nu) Q_n(u,\, v) e^{i\Phi_n(u,\, v)}, \tag{7}$$

where i is the imaginary unit with $i^2 = -1$, and $Q_n(u,\, v)$ is a binary function that is equal to 1 in the points $(u,\, v)$ on the metasurface's plane that are part of a nano-pillar of the n-th type and that is equal to 0 elsewhere.

In order to incorporate the transmission function shown in Eq. (7) into our computational model, we represent both sides of the equation as multi-dimensional discretized numerical arrays. When monochromatic illumination is used (like in this work's case of interest) and we set $\nu = \nu_0$ on both sides of Eq. (7), the terms in the sum on the right-hand side are represented by a 3D array. The first two dimensions represent the spatial dependence on the coordinates $(u,\, v)$, and each pixel represents a different unit cell in the metasurface. Meanwhile, the third dimension represents the different multiplexed nano-pillar arrays indexed by n. As a result, after performing the sum over n to compute $T(u,\, v;\, \nu_0)$, this transmission function is represented by a 2D numerical array, since it depends only on spatial coordinates. Meanwhile, for the broadband illumination case, the numerical arrays representing the functions on both sides of the equation would have one more dimension, which denotes the dependence on frequency ν.

2.4. *Design and characterization pipeline*

At the start of the joint optimization process, the CV task and Attacker models are initialized with pre-loaded parameter values referred to as "Imagenet weights" [37]. Afterwards, both of them

Optics EXPRESS

are independently pre-trained to attain high performance in their respective tasks when receiving the original unobscured dataset images as their inputs, in the absence of the Optics model. Once this initial round of pre-training is completed, the models reach a state which we refer to as "baseline", since it is later on used to initialize the models that are part of the co-training process. The baseline CV task and Attacker models not only serve as benchmarks for the performance of their co-trained counterparts, but their outputs are also used to compute some of the terms in the Optics model's loss function L_{Opt}, as explained in more detail in Supplement 1.

Meanwhile, the Zernike coefficients that parametrized the Optics model's phase profiles are initialized with random values that follow a zero-mean normal distribution. Empirically, this initial state of the Optics model's parameters produces very mild aberrations that do not degrade the quality of the images. This allows the co-trained CV task and Attacker models to follow a smoother evolution of their trainable parameters, where they initially expect high-fidelity inputs and then gradually start adapting to receiving increasingly degraded inputs.

Once the joint optimization process is over, we then proceed to evaluate the trade-off between privacy preservation and CV task performance that is attained by the final state of the co-trained Optics model with its yielded optical aberrations. To do so, we first create a new instance of an Attacker model (referred to as "Independent Attacker" for the rest of this work). The Independent Attacker is initialized with Imagenet weights and is then trained with the inputs degraded by the Optics model's optimized aberrations until it converges to the best performance it can attain under those conditions. Meanwhile, the CV task model starts off with the values that its parameters converged to at the end of the joint optimization process. From this starting point, the CV task model is fine-tuned until it too reaches the best performance it can attain when receiving inputs degraded by the Optics model. In other words, the Optics model's "Privacy-Performance" trade-off is quantitatively characterized by the maximal performance metrics attained by the fine-tuned CV task model and the trained Independent Attacker that receive inputs that have been obscured by the Optics' optimized optical aberrations.

3. Methods

3.1. Image classification models and dataset

Both the CV task and Attacker models used in this work are convolutional neural networks (CNN) with the ResNet50 architecture [38,39], and they each carry out different image classification tasks. The former performs a binary classification task to determine whether the input image is a picture of a person or not. Meanwhile, the latter performs a multi-class classification task to determine what class the object shown in the input image belongs to, with more weight being given to the accuracy in the classification of samples in the non-person classes.

The dataset used in this work was built from the Common Objects in Context (COCO) dataset, which is comprised of high-resolution images of complex everyday scenes, divided into training, validation, and test sets [24]. Only the training and validation datasets have publicly available annotations that indicate what pixel regions belong to each of the object instances in each picture, as well as what class, of the 80 labeled ones, said objects belong to.

In this work, those annotations are used to produce a collection of images that focus only on one object each, by cropping the bounding boxes that contain each labeled object. To produce our datasets, we only kept the cropped regions of interest that had sizes of more than 256 pixels in both width and height, discarding the rest. Afterwards, these regions of interest were resized down to be images with 256 by 256 pixels. Our dataset is divided into training, validation, and test sets, which were constructed to have approximately the same class distribution and ensuring that each of them had at least one sample of each class. The training set contains 35218 images; the validation set, 11334; and the test set, 5913. Of these, 10584 belong to the "person" class in the training set, as is the case for 3402 images in the validation set and 1720 images in the test set.

Optics EXPRESS

Finally, due to the fact that the class label distribution is not balanced for neither the binary classification nor the multi-class classification tasks, we had to carry out the standard practice of using dataset augmentation techniques and incorporating class-dependent weights into the cross-entropy loss functions L_{CV} and L_{Atk} for the training of the classification models to be more robust [7,40].

3.2. Metasurface nano-pillar design

In order to determine the geometrical parameters of the metasurface's nano-pillars to be used for fabrication, FDTD simulations were performed using the MEEP library in the Python programming language [41]. A script was written to simulate the interaction between a plane wave with left circular polarization and a titanium dioxide (TiO_2) nano-pillar placed on top of a fused silica substrate, at normal incidence. In order to compute the field PCE spectra, we first run a single simulation where the nano-pillar is absent, to measure the frequency domain complex amplitude of the original incident field, $E_0(\nu)$. Afterwards, a set of simulations is ran where the nano-pillar is present, to measure the projection into the right circular polarization state of the transmitted field, $E_{\mathrm{T,right}}(\nu)$. With this, the power PCE spectrum is computed as $\mathrm{PCE}(\nu) = \frac{|E_{\mathrm{T,right}}(\nu)|^2}{|E_0(\nu)|^2}$. Meanwhile, the field PCE spectrum was computed as $A(\nu) = \frac{E_{\mathrm{T,right}}(\nu)}{E_0(\nu)}$.

Each of the simulations in this set were ran by using different values for the geometrical parameters of the TiO_2 nano-pillar. After building a library that charts the mapping between different combinations of geometrical parameter values and the yielded PCE spectra, we examined them and selected a set of three combinations that we deemed as the most appropriate for our work. The selection process was primarily guided by the necessity of having distinct aspect ratios in the geometries of the three types of nano-pillars. Additionally, having high polarization conversion rates at different regions of the visible spectrum for the sake of energy efficiency was contemplated too, even though the main experiments in the current work use monochromatic illumination.

3.3. Optical system layout's specifications

The privacy-preserving optical obscurations in this laboratory experiment are produced using the optical system layout described in Section 2.2. The fabricated metasurface producing these obscurations is placed in the Fourier plane of a $4f$ system, while images are projected by a display that is placed in this system's input plane. The obscured pictures are imaged to this $4f$ system's output plane, which coincides with the input plane of a second $4f$ system (with no optical elements in its Fourier plane) whose purpose is to down-scale and relay the obscured images into the sensor of the Basler a2A1920-160ucPRO camera used to capture them. Despite this being a color camera, it was used in this work due to it having better sensitivity and resolution than the monochromatic models that were available. Furthermore, the sensitivity spectra specifications of this camera model had also been used during the simulations performed with color image inputs and broadband illumination, which used the emission spectra specifications of an eMagin SXGA096 OLED micro-display. As shown in Fig. 2 (not to scale), the first $4f$ system is comprised by lenses L3 and L4, whose focal lengths are $f_3 = 150\,\mathrm{mm}$ and $f_4 = 100\,\mathrm{mm}$, respectively. Meanwhile, the second $4f$ system is comprised by lenses L5 and L6, whose focal lengths are $f_5 = 75\,\mathrm{mm}$ and $f_4 = 15\,\mathrm{mm}$, respectively. All lenses used in these $4f$ systems have a diameter of 2 in, with the exception of L6, whose diameter is 0.5 in.

In order for Eq. (1) of our Optics model to hold true in the laboratory experiments, light coming out the first $4f$ system's input plane must be spatially incoherent. To attain this, we use a Texas Instruments DLP Lightcrafter 6500 digital micro-mirror display (DMD) illuminated by a high-power LED to project the input images. Before being incident on the DMD, the light is expanded and collimated by lenses L1 (with $f_1 = 25.4\,\mathrm{mm}$) and L2 (with $f_2 = 200\,\mathrm{mm}$), which

Optics EXPRESS

have diameters of 1 in and 2 in, respectively. Additionally, in order for light incident on the metasurface to be imbued with the intended phase modulation profile, it must have a left-handed circular polarization state. As explained in Section 2.2, the obscured images are encoded by the portion of the light that is imbued with the geometrical phase modulation and that has its polarization's handedness switched by the metasurface's nano-pillar arrays. Thus, it is necessary to filter out the left-handed circular polarization component (the unconverted portion) of the light coming out of the metasurface, and to keep just the right-handed circularly polarized light carrying the obscured image. To attain these required manipulations of light's polarization state, we use linear polarizers and quarter-wave plates.

Imparting the exact quarter-wave retardance is crucial to properly prepare the pure state of left-handed circular polarization. As such, light incident on these wave plates must have the wavelength for which the necessary retardance value is attained. In this laboratory experiment, we used quarter-wave plates that work with 632.8 nm light, which was obtained by spectrally filtering light emitted by the illumination LED with a band-pass filter that is centered at this wavelength with a 1 nm full width at half maximum. It should also be noted that the necessity for exact quarter-wave retardance currently represents a limitation for experimental demonstrations, both in terms of laboratory experiments and real-world deployments, of the proposed system on applications that involve inputs with broadband spectra. As a result, the complementary studies with color image inputs that explore the geometrical phase metasurface's ability to produce independent PSFs for each color channel were performed via computational simulations only, where ideal broadband quarter-wave plates were imposed to be part of the system's configuration.

3.4. Capture acquisition and processing

Since the dataset used in this work contains thousands of images, it is necessary to automate the processes of projecting them and capturing them. These processes are attained by preparing video files with the images of the training, validation and test sets. As explained in Section 2.2, we can opt to project grayscale versions of the dataset images on the DMD, which is the approach we took when producing the videos' frames. The DMD is then used to play these videos at 5 frames per second while the camera is set to capture images at 10 frames per second. With this, two copies of each image are temporarily saved to the computer's memory, but only one of them is ultimately kept.

The projected images underfill the camera's sensor, whether they are unobscured in the absence of the metasurface or obscured in its presence. As such, after the capture acquisition process is complete, the obtained captures need to be processed to crop the region of interest containing the obscured images. Furthermore, since the CV task and Attacker models are designed to take input images with 256 by 256 pixels, the obscured images need to be resized to have these dimensions after the cropping. The final processing step is converting these captured obscured images from color (red) to grayscale, since the downstream CV task and Attacker models used in this work's main laboratory-based experiment are designed to take monochromatic grayscale inputs.

4. Results

4.1. Attained privacy-performance trade-offs

As explained in Section 2.4, the evaluation of the Optics model's Privacy-Performance trade-off is preceded by separate rounds of fine-tuning the parameters of the CV task model and training an Independent Attacker model while the Optics model's parameters are fixed. These routines have the purpose of adjusting the parameters of the CV task and Independent Attacker models to allow them to attain their best possible performance when receiving obscured inputs. Initially, this process is carried out using the digital Optics model, which generates simulations of what the images obscured by the fabricated metasurface's PSF during the laboratory experiment would

Optics EXPRESS

look like. After the laboratory experiment where the obscured dataset images are captured and processed, these routines are carried out again by using these obscured images directly as inputs for the CV task and Independent Attacker models (without needing to use the digital Optics model).

However, in order to properly interpret the performance metrics measured from the above processes, they need to be compared to those of benchmark models that receive unobscured images (in the absence of the metasurface's aberrations) as their inputs. The benchmark CV task model attains an average precision (AP) of 95.7% on its binary classification task of determining which inputs belong to the "person" class. Meanwhile, the benchmark Independent Attacker model attains an accuracy of 75.4% when classifying the inputs that belong to the 79 non-person classes (a performance metric referred to as "non-person accuracy" in this work).

As for the models that take obscured images generated by the Optics model's simulation as inputs, the fine-tuned CV task model attains an AP of 77.1% while the corresponding Independent Attacker model attains a non-person accuracy of 33.5%. On the other hand, the CV task and Independent Attacker models that take the obscured images captured in the laboratory experiment as their inputs attain an AP of 67.0% and a non-person accuracy of 17.5%, respectively. All the reported metrics can be more clearly visualized in the Privacy-Performance trade-off plot shown in Fig. 4(a). For the sake of comparison, the analogous results corresponding to the simulation that studies the broadband illumination case are reported in this plot as well: The benchmark CV task and Independent Attacker models attain an AP of 98.2% and a non-person accuracy of 81.1%, respectively, when taking unobscured color images as inputs. Meanwhile, the fine-tuned CV task model and the trained Independent Attacker that take obscured color inputs attain an AP of 85.9% and a non-person accuracy of 36.4%, respectively.

4.2. Comparison between simulation and laboratory experiments

The results presented in Fig. 4(a) indicate that the optical aberrations yielded by the fabricated metasurface in the laboratory set-up permit a reduced performance for both the CV task and Independent Attacker models, compared to the obscurations produced by the digital Optics model. This is to be expected, given the presence of noise in the laboratory captures, as well as their qualitative differences with respect to the Optics model's simulated obscured images. An even more degraded Independent Attacker performance is favorable, as it translates into better privacy preservation against the classification task it carries out. However, a reduced performance of the CV task model prompted further examination.

A more in-depth comparison between the two CV task models that receive obscured monochromatic inputs was carried out by plotting their Precision-Recall (PR) curves, which are shown in Fig. 4(b). These plots offer a wider perspective of these models' performance, since the AP values reported above are equal to the area under the corresponding PR curve. These plots are built making use of the probability scores that the CV task models assign to each test dataset sample. These scores are the model's predicted probability of each sample belonging to the "person" class. As discussed in more detail in the deep learning algorithm literature [7], the more consistently that a classification model assigns higher scores to the correct classes in its predictions, the better its performance will be. This will ultimately translate into higher precision values, and thus a higher AP metric, as is the case with the benchmark CV task model that takes the original unobscured images as inputs. On the other hand, using the obscured images as inputs causes the CV task models to attain notably lower precision values when compared to the aforementioned benchmark model. However, the gap between the PR curve of the CV task model that takes the laboratory experiment's captures as inputs and that of the CV task model that uses the Optics model's simulated obscured images is not as wide as the gap between the latter PR curve and that of the benchmark model. In other words, the level of performance that was expected from the simulations is mostly retained during the laboratory experiment.

Optics EXPRESS

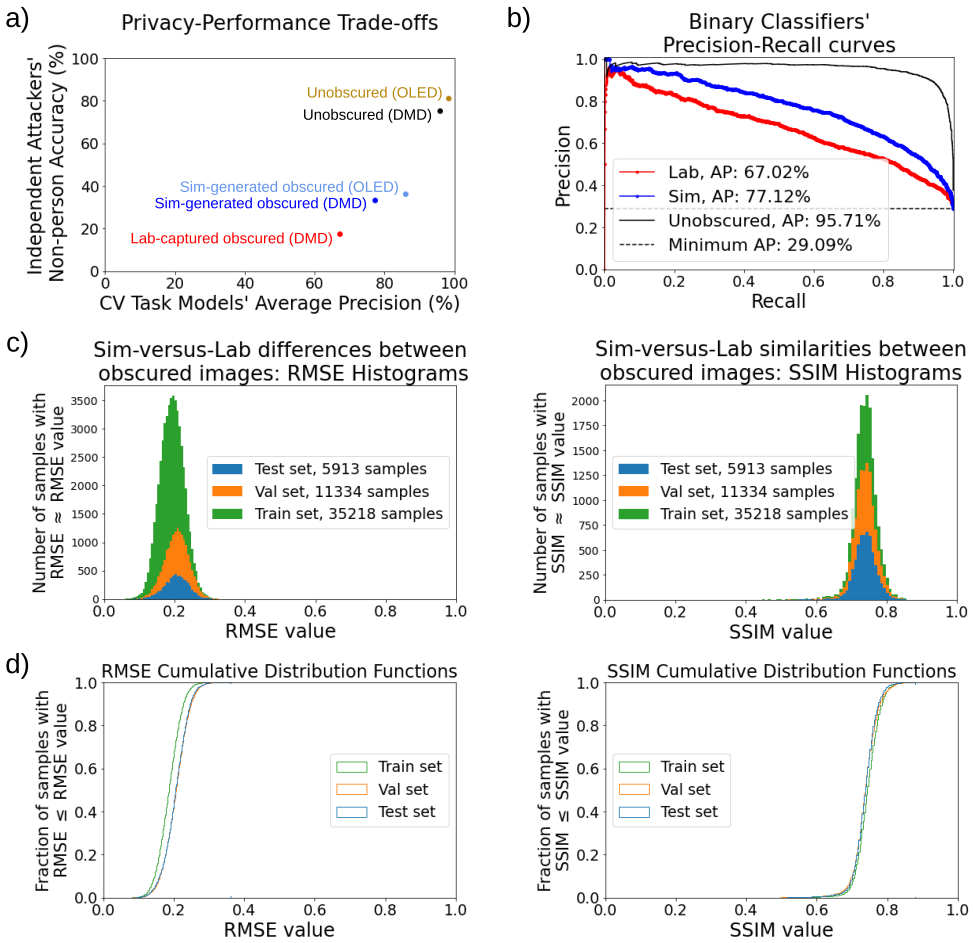

Fig. 4. a) Privacy-Performance trade-offs quantified by the maximal performance metrics of the benchmark models receiving unobscured inputs, as well as those of the models that received obscured inputs obtained either by the Optics model's simulation or the laboratory experiment's captures. b) Precision-Recall (PR) curves of the CV task models that work with monochromatic input images. c) Histograms of the structural similarity index measure (SSIM) and root-mean-square error (RMSE) between the obscured images captured in the laboratory and their corresponding counterparts that had been yielded by the Optics model's simulations. d) Cumulative distribution functions that display the information from the histograms in c) in a more condensed manner.

We finalize our comparison between the main simulation and laboratory experiment by quantitatively measuring the differences and similarities between the obscured images themselves. The former are quantified by the root-mean-squared-error (RMSE) between the normalized obscured images simulated by the Optics model and those obscured by the fabricated metasurface's PSF, captured during the laboratory experiment. Meanwhile, the similarities are quantified by the structural similarity index measure (SSIM) [42] between the two types of obscured images. Histograms and cumulative distribution functions for the measured RMSE and SSIM values of all datasets' samples are shown in Fig. 4(c) and (d). The average RMSE is around 0.2, which is 20% of the dynamic range that the normalized obscured images have. This translates into a relatively high average pixel-wise difference between images obtained during the laboratory

Optics EXPRESS

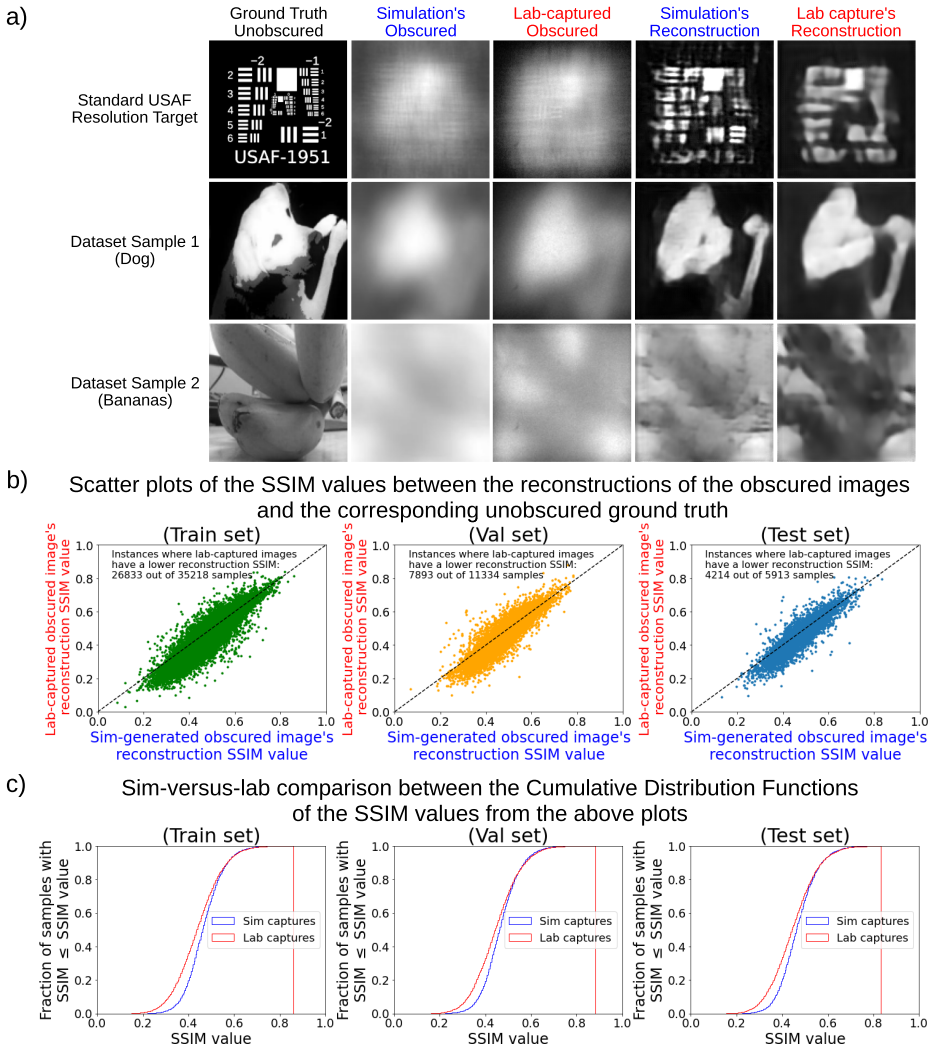

Fig. 5. a) Qualitative comparison between the ground truth unobscured input samples, the corresponding obscured images simulated by the Optics model, the images obscured during the laboratory experiment, as well as the respective image restorations yielded by optimized dedicated image reconstruction models. b) Scatter plots comparing the reconstruction quality of the obscured images obtained in the laboratory (vertical axis) versus that of those simulated by the Optics model (horizontal axis). The dashed lines are the identity function. c) Cumulative distribution functions showing the statistical distribution of the SSIM values plotted in b), but separating the values from each axis into different curves.

and the simulation experiments. Despite this, the distribution of SSIM values has a fairly high mean, of more than 70% of the SSIM function's dynamic range, which translates into a high degree of perceptual similarity between the compared images as a whole. This high degree of global similarity co-existing with relatively high average pixel-wise differences can be explained by observing the qualitative comparison shown in Fig. 5(a). In the displayed examples, it is possible to appreciate that the laboratory experiment's captures are very visually similar to their simulation counterparts, but the amount of contrast is the main difference. The overall shape

Optics EXPRESS

and position of the different bright and dark regions is mostly retained between simulation and laboratory captures, but the latter's images tend to have a higher contrast.

4.3. Robustness against reconstruction attackers

As a complementary avenue to evaluate the optical aberrations' ability to preserve information privacy, we performed a separate analysis where we trained and evaluated the performance of independent image reconstruction deep learning algorithms that attempt to restore the obscured information. It should be noted that the co-trained Optics model in this work was not optimized to fend off against reconstruction attacks. Such an approach would be possible by using one such image reconstruction model as the Attacker in the joint optimization scheme described in Section 2.1. However, demonstrating the usage of this process with that additional type of Attacker algorithm is beyond the scope of the current work.

In the present analysis, we train two separate instances of the same image reconstruction algorithm model. One is trained to reconstruct the obscured images generated by the already optimized Optics computational model, and the other is trained to reconstruct those yielded by the already fabricated metasurface used in the laboratory experiment. The model architecture and its training process are like those presented in [43,44]. This training process involves the usage of a training set consisting of known pairs of obscured images and their corresponding ground truth unobscured counterparts, as well as having knowledge of the PSF that produced the obscurations. Executing this approach to training a privacy-breaching model constitutes a simulacrum of the worst-case scenario in the effort of preserving pre-capture privacy, wherein the attackers get access to the device (or just the pre-capture privacy module) and are able to take the necessary measurements to train their reconstruction model (like the obscured-unobscured image pairs and the system's PSF). With that into consideration, this complementary analysis provides insight about the robustness that the final optimized state of the Optics has against this worst-case situation.

Figure 5(a) offers a qualitative comparison between some example reconstructions yielded by each reconstruction model and the ground truth unobscured version of the samples in question. The general trend is that most of the perceptual information of the images was restored, except for the components with higher spatial frequencies, such as the finer details of textures, written text, background objects, and certain identifying features in living creatures. Furthering this analysis, the quality of these reconstructions was quantified by measuring the SSIM between the reconstructed images and the corresponding ground truth unobscured sample. This was done for both the reconstructions of the obscured images captured in the laboratory experiment and those of the obscured images generated by the Optics model's simulation. Figure 5(b) shows scatter plots where the former's SSIM values are in the vertical axis and the latter's are in the horizontal axis. These plots show a correlation between the quality of the reconstructions of the two types of obscured images: If a given sample had a simulated obscured image with a reconstruction of high SSIM quality, then the corresponding counterpart captured in the laboratory would tend to also have a reconstruction of a high SSIM quality. This is to be expected, given the high similarity between the two types of obscured images, which was discussed in Section 4.2. However, for the overwhelming majority of the samples, the reconstruction quality of the obscured images captured in the laboratory experiment is lower than that of the obscured images simulated by the Optics model. This can be further appreciated in the cumulative distribution function plots shown in Fig. 5(c). For the sets of SSIM values associated with each type of obscured images, there is a noticeable gap between both distributions for the lower percentiles of values, which only starts closing beyond approximately the 70th percentile (vertical axis coordinate). This means that the worse reconstructions have a noticeably lower quality for the laboratory experiment's captures, while the best reconstructions have the same level of quality for both types of obscured images.

Optics EXPRESS

To finalize the analysis, we trained independent instances of the Attacker and CV task models until they converged to their maximal performance on their respective set of reconstructed images (whether from the simulation or the laboratory experiment). After reconstructing the images that had been obscured in simulation by the computational Optics model, the Attacker model could attain a non-person accuracy of 37.4% while the CV task model could attain an AP of 79.5%. Meanwhile, when using the reconstructions of the obscured images captured in the laboratory experiment, the Attacker and the CV task model attained a non-person accuracy of 32.5% and an AP of 76.9%, respectively. These additional results are plotted in Fig. 6(a), where they are compared against those reported in Section 4.1 and Fig. 4(a) for the simulation and laboratory experiment with monochromatic images. Similarly, the PR curves corresponding to the CV task models that were trained with the reconstructed images are plotted in Fig. 6(b), where they are compared against the PR curves that had been presented in Fig. 4(b). In the simulation-generated obscured images' case, neither the Attacker nor the CV task models saw significant performance improvements upon using the corresponding reconstructions, which is contrasted with the evident performance improvements that happened in the laboratory-captured images' case. Finally, the performance metrics that both image classification models attained when using the two types of reconstructed images ended up being very close to each other, which is consistent with the similarities and differences between the distributions of image quality values that had been shown in Fig. 5(c): The higher floor of image quality attained with the simulation-generated obscured images' reconstructions is what allows the image classifier models to achieve a slightly higher performance when using these images compared to the case where they use the reconstructions of the obscured laboratory captures.

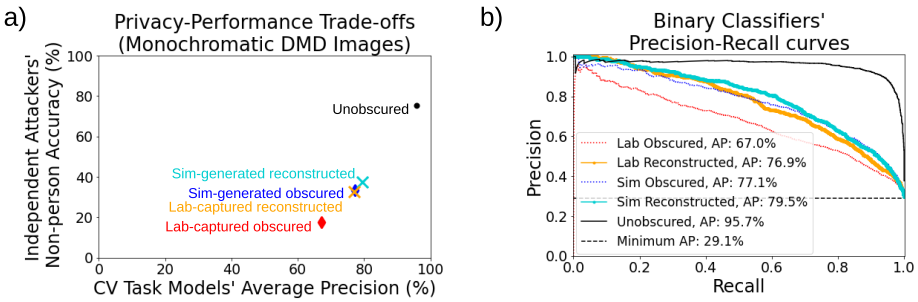

Fig. 6. a) Privacy-Performance trade-offs quantified by the maximal performance metrics of the Attacker and CV task models that take monochromatic images projected by the DMD as inputs, comparing the maximal performance attained when the image classification models use obscured, reconstructed, or unobscured images. b) Precision-Recall (PR) curves of the CV task models whose AP was plotted in a).

5. Discussion and conclusion

In this work, we demonstrated the joint optimization of a pre-capture privacy metasurface-based optical system and a binary classification computer vision task model. The pre-capture privacy optical module does not incur on additional computational overhead for the device carrying out the computer vision task, and it drastically decreases the maximal performance of the privacy-breaching attackers that it was designed to fend off against. This comes at the cost of having the performance of the CV task model also being reduced. However, the trade-off between privacy and performance preservation favors the latter, since the CV task model's performance drop-off is not as drastic as that of the hindered Attacker model.

Furthermore, the optimized optical system displays a decent level of competence at hindering attackers that use advanced image reconstruction algorithms despite having being trained to fend

Optics EXPRESS

off image classification attackers instead. Even though most of the perceptual information can be restored by strong reconstruction models, we demonstrated how some key features with high spatial frequency components were unable to be reconstructed. A future avenue of research is to explore if the robustness against reconstruction attackers could be improved by incorporating them as the adversarial component in the joint optimization process that was demonstrated in this work. However, given the strength of deep-learning-based reconstruction algorithms, attaining a high level of performance in fending off against them would also require to engineer more complex and sophisticated ways to optically degrade the images before they are captured. Tackling the problems associated with building such systems with arrays of metasurfaces that work at visible wavelengths and studying the Privacy-Performance trade-offs that they attain would constitute interesting research paths, but it is currently out of the scope of the present work. Still, we have provided the groundwork to facilitate the inclusion of the technical improvements that are necessary to pursue those research goals using the design and optimization framework demonstrated in this work. Additional technical improvements, such as pursuing miniaturization by using only metasurface-based optical components or resolving the technical problems that limit experimental implementations for applications with broadband illumination, could add further practical value to these future research endeavors.

Funding. University of Rochester (UR).

Acknowledgments. C. Villegas Burgos would like to thank the Humanities, Science and Technology Council of Mexico (Consejo Nacional de Humanidades, Ciencias y Tecnologías, CONAHCYT) for the financial support provided by the fellowship they granted him. We also thank Andrew Berkovich and Reid Pinkham for many discussions.

Disclosures. The authors declare no conflict of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Supplemental document. See Supplement 1 for supporting content.

References

1. A. Acquisti, L. Brandimarte, and J. Hancock, "How privacy-s past may shape its future," Science **375**(6578), 270–272 (2022).
2. F. Pittaluga and S. J. Koppal, "Privacy preserving optics for miniature vision sensors," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), pp. 314–324.
3. M. S. Ryoo, B. Rothrock, C. Fleming, *et al.*, "Privacy-preserving human activity recognition from extreme low resolution," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, (AAAI Press, 2017), AAAI'17, p. 4255–4262.
4. N. Raval, A. Machanavajjhala, and L. P. Cox, "Protecting visual secrets using adversarial nets," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2017), pp. 1329–1332.
5. Z. Wu, Z. Wang, Z. Wang, *et al.*, "Towards privacy-preserving visual recognition via adversarial training: A pilot study," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds. (Springer International Publishing, Cham, 2018), pp. 627–645.
6. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**(7553), 436–444 (2015).
7. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016). http://www.deeplearningbook.org.
8. G. Wetzstein, H. Ikoma, C. Metzler, *et al.*, "Deep optics: Learning cameras and optical computing systems," in *2020 54th Asilomar Conference on Signals, Systems, and Computers*, (2020), pp. 1313–1315.
9. Y. E. Peng, A. Veeraraghavan, W. Heidrich, *et al.*, "Deep optics: Joint design of optics and image recovery algorithms for domain specific cameras," in *ACM SIGGRAPH 2020 Courses*, (Association for Computing Machinery, New York, NY, USA, 2020), SIGGRAPH '20.
10. H. Arguello, J. Bacca, H. Kariyawasam, *et al.*, "Deep Optical Coding Design in Computational Imaging: A data-driven framework," IEEE Signal Process. Mag. **40**(2), 75–88 (2023).
11. J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019).
12. C. A. Metzler, H. Ikoma, Y. Peng, *et al.*, "Deep optics for single-shot high-dynamic-range imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020).
13. Y. Liu, C. Zhang, T. Kou, *et al.*, "End-to-end computational optics with a singlet lens for large depth-of-field imaging," Opt. Express **29**(18), 28530–28548 (2021).
14. Q. Sun, C. Wang, Q. Fu, *et al.*, ACM Trans. Graph. **40** (2021).

Optics EXPRESS

15. E. Nehme, B. Ferdman, L. E. Weiss, *et al.*, "Learning Optimal Wavefront Shaping for Multi-Channel Imaging," IEEE Trans. Pattern Anal. Mach. Intell. **43**(7), 2179–2192 (2021).
16. Z. W. Wang, V. Vineet, F. Pittaluga, *et al.*, "Privacy-preserving action recognition using coded aperture videos," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2019), pp. 1–10.
17. Y. Ishii, S. Sato, and T. Yamashita, "Privacy-aware face recognition with lensless multi-pinhole camera," in *Computer Vision – ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, eds. (Springer International Publishing, Cham, 2020), pp. 476–493.
18. C. Hinojosa, J. C. Niebles, and H. Arguello, "Learning privacy-preserving optics for human pose estimation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), pp. 2553–2562.
19. C. Hinojosa, M. Marquez, H. Arguello, *et al.*, "Privhar: Recognizing human actions from privacy-preserving lens," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, *et al.*, eds., (Springer Nature Switzerland, Cham, 2022), pp. 314–332.
20. J. Lopez, C. Hinojosa, H. Arguello, *et al.*, "Privacy-preserving optics for enhancing protection in face de-identification," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2024), pp. 12120–12129.
21. B. Bai, Y. Luo, T. Gan, *et al.*, "To image, or not to image: class-specific diffractive cameras with all-optical erasure of undesired objects," eLight **2**(1), 14 (2022).
22. B. Bai, H. Wei, X. Yang, *et al.*, "Data-Class-Specific All-Optical Transformations and Encryption," Adv. Mater. **35**(31), 2212091 (2023).
23. C. M. V. Burgos, P. Xiong, L. Qiu, *et al.*, "Co-designed metaoptoelectronic deep learning," Opt. Express **31**(4), 6453–6463 (2023).
24. T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds., (Springer International Publishing, Cham, 2014), pp. 740–755.
25. N. Yu and F. Capasso, "Flat optics with designer metasurfaces," Nat. Mater. **13**(2), 139–150 (2014).
26. D. N. Neshev and A. E. Miroshnichenko, "Enabling smart vision with metasurfaces," Nat. Photonics **17**(1), 26–35 (2023).
27. J. P. Balthasar Mueller, N. A. Rubin, R. C. Devlin, *et al.*, "Metasurface Polarization Optics: Independent Phase Control of Arbitrary Orthogonal States of Polarization," Phys. Rev. Lett. **118**(11), 113901 (2017).
28. B. Wang, F. Dong, Q.-T. Li, *et al.*, "Visible-Frequency Dielectric Metasurfaces for Multiwavelength Achromatic and Highly Dispersive Holograms," Nano Lett. **16**(8), 5235–5240 (2016).
29. C. M. V. Burgos, T. Yang, Y. Zhu, *et al.*, "Design framework for metasurface optics-based convolutional neural networks," Appl. Opt. **60**(15), 4356–4365 (2021).
30. J. W. Goodman, *Introduction to Fourier optics* (Roberts and Company Publishers, 2005), 3rd ed.
31. I. Malkiel, M. Mrejen, A. Nagler, *et al.*, "Plasmonic nanostructure design and characterization via Deep Learning," Light: Sci. Appl. **7**(1), 60 (2018).
32. Z. Liu, D. Zhu, S. P. Rodrigues, *et al.*, "Generative Model for the Inverse Design of Metasurfaces," Nano Lett. **18**(10), 6570–6576 (2018).
33. X. Shi, T. Qiu, J. Wang, *et al.*, "Metasurface inverse design using machine learning approaches," J. Phys. D: Appl. Phys. **53**(27), 275105 (2020).
34. C. Munley, W. Ma, J. E. Fröch, *et al.*, "Inverse-Designed Meta-Optics with Spectral-Spatial Engineered Response to Mimic Color Perception," Adv. Opt. Mater. **10**(20), 2200734 (2022).
35. M. Born, E. Wolf, A. B. Bhatia, *et al.*, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light* (Cambridge University Press, 1999).
36. L. N. Thibos, R. A. Applegate, J. T. Schwiegerling, *et al.*, Journal of Refractive Surgery **18**, (2002).
37. J. Deng, W. Dong, R. Socher, *et al.*, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, (Ieee, 2009), pp. 248–255.
38. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, (2012), pp. 1097–1105.
39. K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), pp. 770–778.
40. M. Abadi, A. Agarwal, P. Barham, *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015). Software available from tensorflow.org.
41. A. F. Oskooi, D. Roundy, M. Ibanescu, *et al.*, "Meep: A flexible free-software package for electromagnetic simulations by the FDTD method," Comput. Phys. Commun. **181**(3), 687–702 (2010).
42. Z. Wang, A. Bovik, H. Sheikh, *et al.*, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. on Image Process. **13**(4), 600–612 (2004).
43. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds. (Springer International Publishing, Cham, 2015), pp. 234–241.
44. J. D. Rego, K. Kulkarni, and S. Jayasuriya, "Robust lensless image reconstruction via psf estimation," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (2021), pp. 403–412.