Supplemental Document

Optics EXPRESS

Co-designed pre-capture privacy optics for computer vision: supplement

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*

Supplement DOI: https://doi.org/10.6084/m9.figshare.28941362

Parent Article DOI: https://doi.org/10.1364/OE.557638

Co-designed pre-capture privacy optics for computer vision: supplemental document

In this Supplemental Document, we show the derivation for the equations that are used at the core of our computational model for the imaging system used in our work. Additionally, we provide additional detailed explanations for the design of the loss function terms that are used in the gradient-based joint optimization process of our models. Furthermore, we expound on the design and implementation of the simulation of the camera's image signal processing (ISP) algorithm that was incorporated into our Optics model. Lastly, we examine the class-specific performance of the image classifier models, to identify which classes of objects are obscured the most by the optimized optical aberrations.

1. LINEAR IMAGING SYSTEMS WITH BROADBAND SPATIALLY INCOHERENT ILLUMINATION AND AN ANISOTROPIC OPTICAL ELEMENT

For a linear imaging system with spatially incoherent monochromatic illumination under the scalar Optics regime, the intensity distribution in the output plane is given by the convolution between the intensity distribution in the input plane and the imaging system's point-spread function (PSF) [1]:

$$I_{\text{out}}(x,\, y) = I_{\text{in}}\left(\tilde{\xi},\, \tilde{\eta}\right) * \text{PSF}\left(\tilde{\xi},\, \tilde{\eta}\right). \tag{S1}$$

In the main document, we based our Optics computational model on a similar equation, for the case of a linear imaging system with spatially incoherent illumination and a wavelength-dependent PSF:

$$\mathcal{G}_{\text{out}}(x,\, y;\, \nu) = \mathcal{G}_{\text{in}}(\tilde{\xi},\, \tilde{\eta};\, \nu) * \text{PSF}\left(\tilde{\xi},\, \tilde{\eta};\, \nu\right) = \iint \mathcal{G}_{\text{in}}(\tilde{\xi},\, \tilde{\eta};\, \nu)\text{PSF}(x-\tilde{\xi},\, y-\tilde{\eta};\, \nu)d\tilde{\xi}d\tilde{\eta}, \tag{S2}$$

where $\mathcal{G}_{\text{in}}$ and $\mathcal{G}_{\text{out}}$ denote the power spectral density distribution at the input and output planes, respectively. In this section, we provide a derivation for this equation. Our analysis also takes into consideration the anisotropic properties of the metasurface's nano-pillars that are present in the system used in our work.

Since we are dealing with an anisotropic optical element, we need to use vector Optics analysis [2, 3]. We start by considering the Jones matrix associated with one of these anisotropic nano-pillars:

$$\widehat{T}(0) = \begin{bmatrix} S_1 e^{i\varphi_1} & 0 \\ 0 & S_2 e^{i\varphi_2} \end{bmatrix}. \tag{S3}$$

Without loss of generality, we assume that light propagates along the Cartesian z axis. Additionally, $S_{1,2}$ and $\varphi_{1,2}$ are the modulation to the magnitude and phase, respectively, of the incident light's electric field's components on the directions aligned with the material's main axes (assumed to coincide with the Cartesian x and y axes) after the light propagated through it. If the nano-pillar now becomes rotated around the propagation axis so that there is an angle θ between the Cartesian axes and the material's main axes, the Jones matrix would then be given by:

$$\begin{aligned} \widehat{T}(\theta) &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} S_1 e^{i\varphi_1} & 0 \\ 0 & S_2 e^{i\varphi_2} \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \\ \widehat{T}(\theta) &= \begin{bmatrix} S_1 e^{i\varphi_1}\cos^2\theta + S_2 e^{i\varphi_2}\sin^2\theta & \left(S_1 e^{i\varphi_1} - S_2 e^{i\varphi_2}\right)\cos\theta\sin\theta \\ \left(S_1 e^{i\varphi_1} - S_2 e^{i\varphi_2}\right)\cos\theta\sin\theta & S_1 e^{i\varphi_1}\sin^2\theta + S_2 e^{i\varphi_2}\cos^2\theta \end{bmatrix}. \end{aligned} \tag{S4}$$

We will now perform the analysis of light's propagation through our imaging system. In this imaging system, a micro-display screen is placed in the front focal plane of a lens with focal length f_1, and our metasurface is placed in the back focal plane of said lens. The metasurface's

plane also coincides with the front focal plane of a second lens, with focal length f_2, and whose back focal plane is the output plane of the overall imaging system. Propagation from the display to the front of the metasurface is functionally the same as the propagation from the back of the metasurface to the imaging system's output plane. Additionally, the field at a point at the back of the metasurface $\mathrm{u}_{m'}$ is given by the field at the front of the metasurface u_m and the Jones matrix for that point:

$$\begin{bmatrix} \mathrm{u}_{m'x}(u,\, v;\, \nu) \\ \mathrm{u}_{m'y}(u,\, v;\, \nu) \end{bmatrix} = \begin{bmatrix} T_{11}(u,\, v;\, \nu) & T_{12}(u,\, v;\, \nu) \\ T_{21}(u,\, v;\, \nu) & T_{22}(u,\, v;\, \nu) \end{bmatrix} \begin{bmatrix} \mathrm{u}_{mx}(u,\, v;\, \nu) \\ \mathrm{u}_{my}(u,\, v;\, \nu) \end{bmatrix}, \tag{S5}$$

where $(u,\, v)$ are the spatial coordinates on the metasurface's plane, and ν denotes the frequency components of the field and the frequency dependence of the Jones matrix's elements.

With the above in mind, we start our analysis by first examining the propagation from the front focal plane of a lens to its back focal plane. The vector and frequency components of the field obey the same wave equation, to which Fresnel propagation is a solution in the paraxial regime. In the absence of anisotropic propagation media, a scalar Optics treatment is sufficient for this part of the analysis. Said treatment has already been used in [1], assuming a thin achromatic lens; for brevity sake, we cite the result. Let u_o be the field of frequency ν at the front focal plane of a lens with focal length f; then the field u_f at the back focal plane of the lens is given by:

$$\mathrm{u}_f(u,\, v;\, \nu) = \frac{\nu e^{i2\pi\nu\frac{(2f)}{c}}}{icf} \mathcal{F}\left\{\mathrm{u}_o(\xi,\, \eta;\, \nu)\right\}\Big|_{\left(\frac{\nu}{cf}u,\, \frac{\nu}{cf}v\right)}, \tag{S6}$$

where c is the speed of light in vacuum, and $\mathcal{F}\{\cdot\}$ denotes the 2D Fourier transform, defined as:

$$\mathcal{F}\left\{g(x,y)\right\}\Big|_{(f_x,\, f_y)} = \iint_{-\infty}^{\infty} g(x,y) \exp\left[-i2\pi\left(f_x x + f_y y\right)\right] dx\, dy, \tag{S7}$$

where $(f_x,\, f_y)$ are the spatial frequency coordinates in the spatial frequency domain.

We now move on to the vector Optics analysis of the propagation from the back of the metasurface to the output plane of our imaging system, which are the front and back focal planes of the system's second lens. As such, since the field's components each obey Fresnel propagation, we apply the result of Eq. (S6) to both of them. Thus, the field u_s at the imaging system's output plane is given in terms of the field at the back of the metasurface $\mathrm{u}_{m'}$ as:

$$\mathrm{u}_{sx}(x,\, y;\, \nu) = \frac{\nu e^{i2\pi\nu\frac{(2f_2)}{c}}}{icf_2} \mathcal{F}\left\{\mathrm{u}_{m'x}(u,\, v;\, \nu)\right\}\Big|_{\left(\frac{\nu}{cf_2}x,\, \frac{\nu}{cf_2}y\right)} \tag{S8a}$$

$$\mathrm{u}_{sy}(x,\, y;\, \nu) = \frac{\nu e^{i2\pi\nu\frac{(2f_2)}{c}}}{icf_2} \mathcal{F}\left\{\mathrm{u}_{m'y}(u,\, v;\, \nu)\right\}\Big|_{\left(\frac{\nu}{cf_2}x,\, \frac{\nu}{cf_2}y\right)}, \tag{S8b}$$

where $\mathrm{u}_{m'}$ itself is given in terms of u_m by Eq. (S5).

Without loss of generality, let's assume that right-hand circularly polarized light is incident on the metasurface. In that case, we have:

$$\begin{bmatrix} \mathrm{u}_{mx}(u,\, v;\, \nu) \\ \mathrm{u}_{my}(u,\, v;\, \nu) \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix} \mathrm{u}_m(u,\, v;\, \nu), \tag{S9}$$

where $\mathrm{u}_m(u,\, v;\, \nu)$ is a scalar amplitude that obeys Fresnel propagation. With that, substituting Eq. (S5) and Eq. (S9) into Eq. (S8), we get:

$$\mathrm{u}_{sx}(x,\, y;\, \nu) = \frac{\nu e^{i2\pi\nu\frac{(2f_2)}{c}}}{icf_2} \mathcal{F}\left\{\frac{1}{\sqrt{2}}\left(T_{11}(u,\, v;\, \nu) - iT_{12}(u,\, v;\, \nu)\right)\mathrm{u}_m(u,\, v;\, \nu)\right\}\Big|_{\left(\frac{\nu}{cf_2}x,\, \frac{\nu}{cf_2}y\right)} \tag{S10a}$$

$$\mathrm{u}_{sy}(x,\, y;\, \nu) = \frac{\nu e^{i2\pi\nu\frac{(2f_2)}{c}}}{icf_2} \mathcal{F}\left\{\frac{1}{\sqrt{2}}\left(T_{21}(u,\, v;\, \nu) - iT_{22}(u,\, v;\, \nu)\right)\mathrm{u}_m(u,\, v;\, \nu)\right\}\Big|_{\left(\frac{\nu}{cf_2}x,\, \frac{\nu}{cf_2}y\right)}. \tag{S10b}$$

Similarly, applying the result of Eq. (S6), we have that the field at the front of the metasurface u_m is given in terms of the field at the display plane u_o by:

$$\mathrm{u}_m(u,\, v;\, \nu) = \frac{\nu e^{i2\pi\nu\frac{(2f_1)}{c}}}{icf_1}\mathcal{F}\left\{\mathrm{u}_o(\xi,\, \eta;\, \nu)\right\}\Big|_{\left(\frac{\nu}{cf_1}u,\, \frac{\nu}{cf_1}v\right)}. \tag{S11}$$

We now consider the projection u_{sp} of u_s into the base vector of the left-hand circular polarization state, which has the opposite handedness to that of the light that was incident on the metasurface. For that, we take the complex inner product between u_s and the unit vector $\frac{1}{\sqrt{2}}\begin{bmatrix}1\\ i\end{bmatrix}$, so we have:

$$\mathrm{u}_{sp} = \tfrac{1}{\sqrt{2}}\mathrm{u}_{sx} - \tfrac{i}{\sqrt{2}}\mathrm{u}_{sy}. \tag{S12}$$

Finally, we substitute the values of the Jones matrix from Eq. (S4) into Eq. (S10), and substitute both Eq. (S10) and Eq. (S11) into Eq. (S12). After some work, we get to this expression for the projection of the field at the imaging system's output plane u_{sp} in terms of the field at the input plane:

$$\mathrm{u}_{sp}(x,\, y;\, \nu) = -\frac{\nu^2 e^{i2\pi\nu\frac{(2f_1+2f_2)}{c}}}{c^2 f_1 f_2}\left(\frac{f_1}{f_2}\right)^2 \iint_{-\infty}^{\infty} d\tilde{\xi}\, d\tilde{\eta}\, \mathrm{u}_o\left(-\tfrac{f_1}{f_2}\tilde{\xi},\, -\tfrac{f_1}{f_2}\tilde{\eta};\, \nu\right) h\left(x-\tilde{\xi},\, y-\tilde{\eta};\, \nu\right), \tag{S13}$$

where h is a function given by:

$$h\left(x-\tilde{\xi},\, y-\tilde{\eta};,\, \nu\right) = \mathcal{F}\left\{e^{-i2\theta(u,v;\nu)}\left(\frac{S_1(u,\, v;\, \nu)e^{i\varphi_1(u,v;\nu)} - S_2(u,\, v;\, \nu)e^{i\varphi_2(u,v;\nu)}}{2}\right)\right\}\Big|_{\left(\frac{\nu}{cf_2}\left(x-\tilde{\xi}\right),\, \frac{\nu}{cf_2}(y-\tilde{\eta})\right)}. \tag{S14}$$

With this result, we can find the power spectral density $\mathcal{G}_{sp}$ of this field's projection by as:

$$\begin{aligned}\mathcal{G}_{sp}(x,\, y;\, \nu) &= \lim_{\mathrm{T}\to\infty}\frac{1}{\mathrm{T}}\left\langle\left|\mathrm{u}_{sp}(x,\, y;\, \nu)\right|^2\right\rangle \\ &= \left(\frac{\nu}{cf}\right)^4 \iiiint_{-\infty}^{\infty}\mathcal{G}_o\left(\xi_1,\, \eta_1;\, \xi_2,\, \eta_2;\, \nu\right) \\ &\qquad h\left(x-\xi_1,\, y-\eta_1;,\, \nu\right) h^*\left(x-\xi_2,\, y-\eta_2;\, \nu\right) d\xi_1 d\eta_1 d\xi_2 d\eta_2,\end{aligned} \tag{S15}$$

where $\mathcal{G}_o$ is the power spectral density of the field at the input plane, and we used $f_1 = f_2 = f$ for simplicity. Since light emitted by the display is spatially incoherent, we have $\mathcal{G}_o\left(\xi_1,\, \eta_1;\, \xi_2,\, \eta_2;\, \nu\right) = \kappa\,\mathcal{G}_o\left(\xi_1,\, \eta_1;\, \nu\right)\delta\left(\xi_1-\xi_2,\, \eta_1-\eta_2\right)$, for some constant κ. As such, substituting that definition into Eq. (S15), we at last have:

$$\mathcal{G}_{sp}(x,\, y;\, \nu) = \kappa\left(\frac{\nu}{cf}\right)^4 \iint_{-\infty}^{\infty}\mathcal{G}_o\left(\xi,\, \eta;\, \nu\right)\left|h\left(x-\xi,\, y-\eta;\, \nu\right)\right|^2 d\xi d\eta. \tag{S16}$$

After grouping some scalar factors together and doing some notation changes, Eq. (S13), Eq. (S14), and Eq. (S16) conclude the derivation for the equations used in the main document as part of our Optics model.

2. OPTICS MODEL LOSS FUNCTION

As mentioned in the main document, the loss function L_{Opt} that is used to train the Optics model incorporates the losses L_{CV} and L_{Atk} that are used to train the CV task and Attacker models, to create a feedback loop that couples the optimization processes of the three models. In this work, the CV task and Attacker models are image classification models with a deep learning architecture known as convolutional neural network (CNN) [4, 5]. Since both models carry out classification tasks, both L_{CV} and L_{Atk} are the cross-entropy function, which is commonly used to train classification models. A lower value for the cross-entropy loss function indicates that the model in question produces more accurate classifications with more certainty [6]. There are some distinctions between the functional forms of L_{CV} and L_{Atk}, which stem from the fact that

the CV task model carries out a binary classification task while the Attacker model performs multi-class classification. Furthermore, each set of classes (whether the two in the CV task model's binary classification or the 80 in the Attacker's multi-class classification task) is assigned with a set of scalar weights that are introduced as standard practice to compensate for the imbalanced distribution of samples across the classes [6].

Apart from the L_{CV} and L_{Atk} terms, the Optics model's loss function contains three other terms. The first term is the structural similarity index measure (SSIM) [7], which compares the input image that is passed down into the Optics model with the obscured image that is returned as its output. When minimizing L_{Opt}, we want to minimize this SSIM, so that the obscured image stops resembling the input image. The SSIM term is completely independent of the CV task and Attacker models, unlike the last two terms present in L_{Opt}.

These two remaining terms seek to drive the obscured images returned by the Optics model into producing specific kinds of responses from the CV task and Attacker models. This is done by either minimizing or maximizing the difference between given target outputs and the outputs returned by the convolution layers present in these models' architectures. These target outputs are generated by using separate, fixed (non-trainable) instances of the CV task and Attacker models' architectures that are set to their "baseline" states. As explained in the main document, the CV task and Attacker models are initially pre-trained to attain high performance when receiving the original unobscured images as their inputs. The states that these models reach after this pre-training process are used to initialize the co-trained models; as such, we refer to the model instances that are fixed in these states as "baseline models". The outputs from the baseline models' intermediate convolutional layers provide insight about the features from the input image that the models extract and use to perform their classification tasks [4, 5].

Because of the above, adjusting the Optics model so that the responses from the co-trained classification models to the produced obscured images, $\text{Opt}(I_{\text{in}})$, resemble the responses from the baseline models to the original unobscured images, I_{in}, would allow the former to emulate the latter to attain a high classification performance despite receiving inputs with degraded image quality. The first of the remaining terms works with the models' feature maps, which are the final outputs of their series of convolution layers and that can be interpreted as the model's "perception" to the inputs' features [8]. As such, this term is referred to as perceptual loss . In this work, it seeks to make the feature maps from the co-trained CV task model resemble those from the baseline model as much as possible, while making the feature maps from the co-trained Attacker model resemble random noise, hindering its performance:

$$\begin{aligned} L_{\text{perc}} = E\left[\left|\text{CV}_{\text{cotrained}}^{(\text{Conv layers})}\left(\text{Opt}(I_{\text{in}})\right) - \text{CV}_{\text{baseline}}^{(\text{Conv layers})}(I_{\text{in}})\right|^2\right] \\ + E\left[\left|\text{Atk}_{\text{cotrained}}^{(\text{Conv layers})}\left(\text{Opt}(I_{\text{in}})\right) - \text{ReLU}\left(\text{Random}_{\text{normal}}\right)\right|^2\right], \end{aligned} \tag{S17}$$

where $E[\cdot]$ denotes taking the average over all the elements in the numerical array in between the square brackets, and ReLU is the rectifying linear function that maps positive numbers to themselves and non-positive numbers to 0.

Meanwhile, the final term works with the Fourier transform of the outputs of the models' first convolutional layers, and is referred to as the spatial frequency domain loss. The rationale for its inclusion is that the spatial frequency content of the first layer's outputs determines the feature content in the last layer's outputs. This is because both the Optics model's PSF and the classifier model's first layer's convolution filters modulate the spatial frequency content in the input image before it is passed down to the next layers to extract features. The goal of this term is to minimize the differences in the spatial frequency domain between the outputs of the baseline and co-trained versions of the CV task model, while maximizing said differences for the two versions of the Attacker model:

$$\begin{aligned} L_{\text{freq}} = E\left[\left|\mathcal{F}\left\{\text{CV}_{\text{cotrained}}^{(1\text{st layer})}\left(\text{Opt}(I_{\text{in}})\right)\right\} - \mathcal{F}\left\{\text{CV}_{\text{baseline}}^{(1\text{st layer})}(I_{\text{in}})\right\}\right|^2\right] \\ - E\left[\left|\mathcal{F}\left\{\text{Atk}_{\text{cotrained}}^{(1\text{st layer})}\left(\text{Opt}(I_{\text{in}})\right)\right\} - \mathcal{F}\left\{\text{Atk}_{\text{baseline}}^{(1\text{st layer})}(I_{\text{in}})\right\}\right|^2\right]. \end{aligned} \tag{S18}$$

Finally, it should be noted that both L_{perc} and L_{freq} are also incorporated into the training of the CV task model, making it easier to maintain a similar perception and performance as its baseline counterpart. With this, the parameters α from the Optics model, along with the parameters

β_{CV} and β_{Atk} of the CV task and Attacker models are optimized via gradient descent with the following update rules:

$$\alpha \leftarrow \alpha - \nabla_{\alpha}\left(\lambda_{\text{Opt}}L_{\text{SSIM}} + \lambda_{\text{CV}}L_{\text{CV}} - \lambda_{\text{Atk}}L_{\text{Atk}} + \lambda_{\text{perc}}L_{\text{perc}} + \lambda_{\text{freq}}L_{\text{freq}}\right), \tag{S19a}$$

$$\beta_{\text{CV}} \leftarrow \beta_{\text{CV}} - \nabla_{\beta_{\text{CV}}}\left(L_{\text{CV}} + \lambda_{\text{perc}'}L_{\text{perc}} + \lambda_{\text{freq}'}L_{\text{freq}}\right), \tag{S19b}$$

$$\beta_{\text{Atk}} \leftarrow \beta_{\text{Atk}} - \nabla_{\beta_{\text{Atk}}}\left(L_{\text{Atk}}\right), \tag{S19c}$$

where we use the notation $\nabla_X(\cdot)$ to represent the gradient of the function in parentheses with respect to the variable numerical array X in the subindex, and $\lambda_{\{\cdot\}}$ are scalar weighting factors. Through trial and error, we empirically found that using $\lambda_{\text{Opt}} = 0.7$, $\lambda_{\text{CV}} = 0.2$, $\lambda_{\text{Atk}} = 0.1$, $\lambda_{\text{perc}} = \lambda_{\text{perc}'} = 0.1$, and $\lambda_{\text{freq}} = \lambda_{\text{freq}'} = 0.01$ in this work led to a more stable joint optimization process while also yielding more favorable Privacy-Performance trade-off results.

3. SIMULATED ISP IMPLEMENTATION

From the main document, we have that the raw signal intensity produced by the camera pixels with the k-th type of color filter is given by:

$$I'_{\text{out},k}(x, y) = \int_0^{\infty} d\nu\, R_k(\nu)\left(\text{PSF}\left(\tilde{\xi}, \tilde{\eta}; \nu\right) * \sum_c S_c(\nu) I_{\text{in},c}\left(\tilde{\xi}, \tilde{\eta}\right)\right), \tag{S20}$$

where $R_k(\nu)$ is the sensitivity spectrum of this k-th type of camera pixel, $S_c(\nu)$ is the emission (or reflection) spectrum associated with the c-th color channel of the images $I_{\text{in},c}\left(\tilde{\xi}, \tilde{\eta}\right)$ that are projected by the display into the imaging system that has a frequency-dependent point-spread function $\text{PSF}\left(\tilde{\xi}, \tilde{\eta}; \nu\right)$.

Let's consider the ideal case where the imaging system has unit magnification and is aberration-free. In that case, the PSF would be modeled as a delta function for all frequencies. As such, the camera's raw output signal intensity would be given by:

$$I'_{\text{out},k}(x, y) = \sum_c M_{kc} I_{\text{in},c}\left(x, y\right), \tag{S21}$$

where M_{kc} are the elements of a 3×3 matrix $\bar{\mathbb{M}}$, given by:

$$M_{kc} = \int_0^{\infty} R_k(\nu) S_c(\nu) d\nu. \tag{S22}$$

As such, Eq. (S21) can be viewed as a matrix multiplication between a 3×3 matrix $\bar{\mathbb{M}}$ and a $3 \times N$ matrix $\bar{\mathbb{I}}_{\text{in}}$ representing the projected image, where N denotes the total number of pixels in the digital image. The result of this matrix multiplication would be a $3 \times N$ matrix $\bar{\mathbb{I}}'_{\text{out}}$ representing the camera's raw output signal intensity. This can be written in matrix form as:

$$\bar{\mathbb{I}}'_{\text{out}} = \bar{\mathbb{M}}\bar{\mathbb{I}}_{\text{in}}. \tag{S23}$$

The above examination of the ideal aberration-free case paves the path towards a simple way of simulating the camera's image signal processing (ISP) algorithm, which maps the raw camera signal intensity $\bar{\mathbb{I}}'_{\text{out}}$ to the output digital image $\bar{\mathbb{I}}_{\text{out}}$ in the general case. In this work, given the knowledge of the spectra $R_k(\nu)$ and $S_c(\nu)$, we first compute the constant matrix $\bar{\mathbb{M}}$ whose elements are given by Eq. (S22). After simulating the the camera's raw output signal intensity $\bar{\mathbb{I}}'_{\text{out}}$ with the Optics model using Eq. (S20), we simulate the camera's ISP by using the mapping

$$\textbf{ISP}\left(\bar{\mathbb{I}}'_{\text{out}}\right) = \bar{\mathbb{M}}^{-1}\bar{\mathbb{I}}'_{\text{out}} \tag{S24}$$

to get the simulated output digital image $\bar{\mathbb{I}}_{\text{out}}$ that is later passed as the input to the CV task and Attacker computational models in the system. As such, the simulated camera ISP given by Eq. (S24) is designed to produce output digital images with identical RGB values to those of the system's digital inputs $\bar{\mathbb{I}}_{\text{in}}$ under the ideal circumstances where Eq. (S23) would hold true. That is, the simulated Optics model would become an identity function in the absence of the metasurface's light manipulation and optical aberrations.

There are a couple of additional considerations regarding the implementation of this simulated ISP. As stated in the main document, this work made use of a digital micro-mirror display (DMD)

illuminated by monochromatic illumination to project the images. As such, images produced by an Optics model that simulates the main laboratory experiment's conditions would be expected to yield a monochromatic output that is passed down as the CV task and Attacker models' inputs. As such, the Optics model's final output $\bar{\mathbb{I}}_{out}$ is obtained by taking the average over the color channels of the image $\mathbf{ISP}\left(\bar{\mathbb{I}}'_{out}\right)$ computed using the expression from Eq. (S24). As a minute side-note, $\bar{\mathbb{I}}_{out}$ is still represented as a 3-channel RGB image (where the three color channels are equal to the aforementioned average), because that is the format that the ResNet50 layers [4, 5] in the CV task and Attacker models' architectures expect to receive as inputs. (This architecture could not be adjusted to take single-channel image inputs because that would have not allowed us to use the pre-loaded "Imagenet weights" [9], which were used to initialize the models' parameters, as explained in the main document).

The second consideration has to do with the computation of the constant matrix $\bar{\mathbb{M}}$. When introducing the data of the $S_c(\nu)$ associated with the monochromatic display into Eq. (S22), the resulting matrix is not invertible. To get around this issue, we used the $S_c(\nu)$ data of an eMagin SXGA096 OLED micro-display instead. In other words, the matrix $\bar{\mathbb{M}}$ in Eq. (S24) was computed using the OLED display's $S_c(\nu)$ data, while the image $\bar{\mathbb{I}}'_{out}$ was obtained from Eq. (S20) using the monochromatic display's $S_c(\nu)$ data. The resulting images $\mathbf{ISP}\left(\bar{\mathbb{I}}'_{out}\right)$ have the red color that would be expected from the monochromatic illumination at 632.8 nm used in this work.

4. CO-TRAINED IMAGE CLASSIFICATION MODELS' CLASS-SPECIFIC PERFORMANCE

We present a complementary analysis to study the class-specific performance metrics of the CV task and Attacker models from the simulations and laboratory experiments that work with monochromatic images, while continuing the comparison between both. We start by listing the 80 different classes that the dataset images can belong to, which are grouped together into 12 categories in the annotation data from the source COCO dataset [10]. In indexing order, the classes (listed inside parentheses) belonging to each category (written between quotation marks) are: "Person" (person), "vehicle" (bicycle, car, motorcycle, airplane, bus, train, truck, boat), "outdoor" (traffic light, fire hydrant, stop sign, parking meter, bench), "animal" (bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe), "accessory" (backpack, umbrella, handbag, tie, suitcase), "sports" (flying disc, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket), "kitchen" (bottle, wine glass, cup, fork, knife, spoon, bowl), "food" (banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake), "furniture" (chair, couch, potted plant, bed, dining table, toilet), "electronic" (television, laptop, computer mouse, remote control, keyboard, cellphone), "appliance" (microwave oven, stove oven, toaster, sink, refrigerator), "indoor" (book, clock, vase, scissors, teddy bear, hair drier, toothbrush). It must be clarified that only the 80 individual class labels played a role during the training and evaluation of the models in this work, but the 12 broad categories will be used in this section to simplify the presented data visualization, analysis and discussion.

The purpose of this analysis is to quantitatively determine which types of objects were successfully obscured from the Attacker models and which were still recognizable by them. To that effect, we use the Attacker models' test set predictions (outputs with probability scores for each class returned by a multiclass classifier) to compute the average precision (AP) for each class. As discussed in the main document, a higher class-specific AP value is indicative of a classifier model's ability to consistently distinguish between the "positive samples" (samples belonging to the class of interest) and the "negative samples" (samples belonging to all other classes, in the context of multiclass classification), assigning high probability scores to the former and low ones to the latter. In Fig. S1 a) we present scatter plots whose horizontal axes represent the AP values attained by the benchmark Attacker (trained with unobscured images) for each class, while the vertical axes represent the AP values attained by the Independent Attackers that are trained with obscured images produced by the simulation with the optimized computational Optics model (left plot) or by the metasurface-based optical system in the laboratory experiment (right plot). As such, these plots display data points corresponding to each the 79 non-person classes, using a different type of marker for each of the 11 corresponding categories.

It can be observed that among the classes that the benchmark Attacker could correctly identify with high AP values, most of them can no longer be correctly classified with high AP values by the Independent Attackers that are trained with obscured images (both from the simulation and the laboratory experiment). However, in the case of the simulation-generated obscured images, the corresponding Independent Attacker could retain relatively high AP values for a small amount of classes, mostly from the animal, vehicle and furniture categories. The highest values correspond

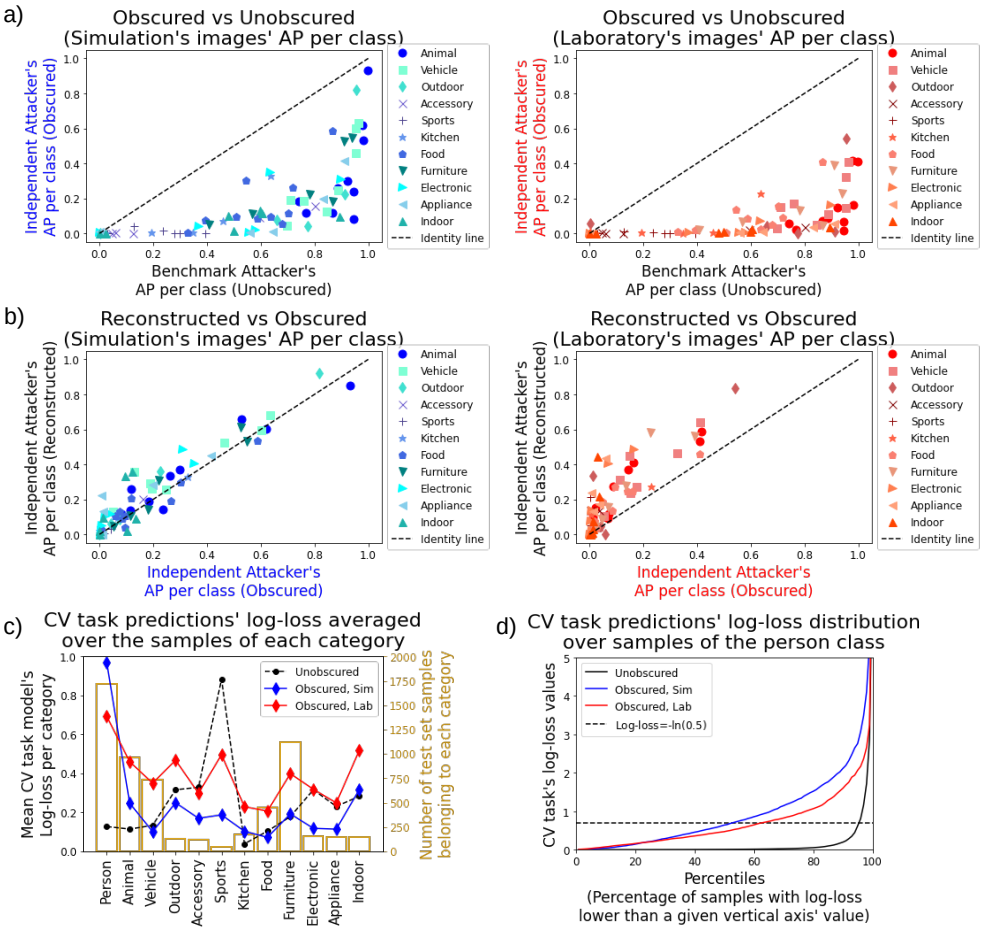

Fig. S1. a) Average precision (AP) values for each non-person class attained by the Independent Attackers that receive obscured inputs, plotted against those attained by the benchmark Attacker that works with unobscured images. b) Class-specific AP values attained by Independent Attackers trained with reconstructed images, plotted against those attained when training with obscured images. c) Averages of the CV task models' predictions' log-loss values over the test set samples belonging to each category. d) Percentile plot of the distribution of the CV task models' predictions' log-loss values over the test set samples that belong to the person class.

to the zebra class (animal category) and the stop sign class (outdoor category), which retain AP values over 77.1% (the AP attained by the CV task model in its binary classification task with simulated obscured images). Meanwhile, in the case of the obscured images captured from the laboratory experiment, the highest class-specific AP value that was attained by the corresponding Independent Attacker was 54.1% for the stop sign class, which is significantly lower than the AP of 67.0% attained by the CV task model that performed the binary classification task of identifying the person class on the same set of obscured images. As such, the Privacy-Performance trade-off of the laboratory experiment's optical obscurations is still shown to favor the CV task model from the perspective of this class-based analysis, since the Independent Attacker model suffers from significant drop-offs in the class-specific AP for all the classes for which the benchmark Attacker had initially attained high performance.

The rest of this section will focus on identifying cases where a favorable Privacy-Performance trade-off fails to be achieved on a per-class basis. We first examine the class-specific performance that an Independent Attacker could attain after being trained with images reconstructed by a dedicated image reconstruction network. As mentioned in the main document, this assumes a scenario where attackers get access to the privacy-preserving optical system and are able to measure its PSF and take captures of obscured-unobscured image pairs to train an image

reconstruction network. Privacy would become compromised in such a scenario, because a trained image reconstruction network allows the Attackers to restore the visual quality of images that had been captured by the device in the past. After training two dedicated image reconstruction networks that restore the quality of the obscured images (one for those from the simulation and one for those from the laboratory experiment), separate Independent Attackers were trained with the corresponding reconstructions of the dataset images. The scatter plots in Fig. S1 b) compare the class-specific AP values that can be attained by the Independent Attackers trained with reconstructed images versus those that were attained by the Independent Attackers that were trained with obscured images. The former values are represented on the vertical axis, while the latter are represented on the horizontal axis. In the case of the obscured images produced by the computational Optics model's simulation, the Independent Attacker that uses their corresponding reconstructions trades off the AP between some classes, leading to higher AP values in some of them at the expense of lower ones in a few others, when compared to the class-specific performance of the counterpart that used the obscured images. However, the Independent Attacker using reconstructed images is favored by this trade-off, since it attains a slightly higher classification accuracy among the non-person classes (37.4% with reconstructed images versus 33.5% with obscured ones). Meanwhile, in the case of the obscured laboratory experiment's captures, using the reconstructed images leads to noticeable improvements in the AP of almost all classes when compared to what was attained with the obscured images. The highest class-specific AP values with the reconstructed laboratory captures correspond to the stop sign class (83.5%) and the bus class (64.4%), which are higher than (or close to) the AP of 67.0% that the CV task model attains on its binary classification when working with the obscured laboratory captures (rather than their reconstructions). Thus, the CV task model is still favored in general from the perspective of this analysis (with the exception of one class), despite not being aided by reconstructed images like the separate Independent Attacker that is.

Lastly, we examine the instances where the CV task model performs poorly. For simplicity of visualization, we limit the analysis to computing the category-specific mean log-loss the CV task model, instead of computing these means for each of the individual 80 classes. That is, we take the sum of all the log-loss values associated with the samples that belong to each category, and then divide it by the total number of samples that belong to that category. The log-loss is defined as the negative natural logarithm, i.e. $-\ln(\cdot)$, of the probability score that the classifier model assigns to the correct class label of a given sample. As mentioned in Section 2, the log-loss quantifies the model's consistency at distinguishing between the different classes; higher probabilities being assigned to the samples' correct labels leads to lower loss values [6]. Additionally, the loss function L_{CV} from Eq. (S19) incorporates scalar weights that are assigned to each class to multiply the result of the $-\ln(\cdot)$ function. For this analysis, we omit the scalar weights that were assigned to the dichotomic positive and negative classes, and focus only on the $-\ln(\cdot)$ function (which is what "log-loss" will refer to). The plot in Fig. S1 c) displays the category-specific mean log-loss values of the CV task models that use obscured images (both in the simulation and the laboratory experiment) for each class, and compares them with those attained by the benchmark CV task model (which used the unobscured images). A histogram of showing the distribution of samples across categories is also shown. It can be observed that the benchmark CV task model (which was trained with unobscured images) has the highest category-specific mean log-loss for the "sports" category, which is the one with the fewest samples. Additionally, both CV task models that work with obscured images have their highest mean log-loss values be those corresponding to the person class. It can also be observed that the CV task model that works with obscured laboratory captures has a lower person-class log-loss value than the one that works with the simulation-generated obscured images, while the latter model attains lower mean log-loss values for all the non-person categories than the former model. However, the high mean log-loss values associated with the class of interest requires further examination. For the simulated obscured images' CV task model, the global mean log-loss (averaged across all test set samples) is 0.4044, whereas the mean log-loss for the person class is 0.9708. Meanwhile, the values of the laboratory captures' CV task model's global mean log-loss and person-class mean log-loss are 0.4661 and 0.6934, respectively. The contrast between the global and person-class mean values comes as a result of the former being driven down by the lower log-loss values of the CV task models' predictions for the non-person samples (which in total comprise two thirds of the test set). Lastly, we examine the distribution of the CV models' log-loss values associated with their predictions for the samples that belong to the person class. To do so, we present the corresponding percentile plots in Fig. S1 d), along with a reference horizontal dashed line for the loss value of $-\ln(0.5) = 0.6931$, which is

the value associated with samples who get assigned with a probability score of 50% for the correct class. In the context of binary classification, positive-class samples that have a log-loss below the 0.6931 threshold would have assigned a probability score of more than 50% to the positive class, resulting in a true positive prediction. From the plots in Fig. S1 d), it can be observed that this is the case with 95% of the unobscured person-class samples classified by the benchmark CV task model. However, upon introducing the optical obscurations, this percentage drops to 52% for the CV task model that works with simulated obscured images, and to 62% for the one that works with the obscured laboratory captures. From this analysis, it can be concluded that the main source of performance degradation in the CV task model is the loss values associated with the class of interest (the person class). The CV task model that works with the obscured laboratory captures is less affected by this problem than its counterpart that worked with the simulation's obscured images. As such, performance improvements (without relying on changing the network architecture) could be pursued with additional rounds of fine-tuning the models' parameters with a more specialized loss function for optimization. For instance, an alternate loss function L'_{CV} with different class-dependent weights focused on prioritizing the positive class performance could be used during this additional fine-tuning instead of the L_{CV} that had been used during co-training (whose weights had the purpose of compensating for the imbalance between the positive and negative classes, as mentioned in Section 2). Furthermore, other specialized loss terms besides L'_{CV} could be introduced during the fine-tuning process, but investigating those possibilities is beyond this work's scope. Regardless, the analysis presented in this section showcases that avenues for improvement can be identified by following performance examinations that are relevant for the applications at hand. Similar procedures would have to be followed if the CV task and Attacker models performed different computer vision tasks or used a different neural network architecture.

REFERENCES

1. J. W. Goodman, *Introduction to Fourier optics* (Roberts and Company Publishers, 2005), 3rd ed.
2. J. W. Goodman, *Statistical Optics*, Wiley Series in Pure and Applied Optics (John Wiley & Sons, Nashville, TN, 2015), 2nd ed.
3. E. Wolf, *Introduction to the theory of coherence and polarization of light* (Cambridge University Press, Cambridge, England, 2007).
4. 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.
5. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* (2016), pp. 770–778.
6. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016). http://www.deeplearningbook.org.
7. Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, IEEE Transactions on Image Process. **13**, 600 (2004).
8. J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016,* B. Leibe, J. Matas, N. Sebe, and M. Welling, eds. (Springer International Publishing, Cham, 2016), pp. 694–711.
9. 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.
10. 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.