Learning a Dynamic Privacy-preserving Camera Robust to Inversion Attacks

Jiacheng Cheng[1], Xiang Dai[1], Jia Wan[2], Nick Antipa[1], and Nuno Vasconcelos[1]

[1] University of California, San Diego
[2] Harbin Institute of Technology, Shenzhen
{jicheng,xidai,nantipa,nvasconcelos}@ucsd.edu, jiawan1998@gmail.com

Abstract. The problem of designing a privacy-preserving camera (PPC) is considered. Previous designs rely on a static point spread function (PSF), optimized to prevent detection of private visual information, such as recognizable facial features. However, the PSF can be easily recovered by measuring the camera response to a point light source, making these cameras vulnerable to PSF inversion attacks. A new dynamic privacy-preserving (DyPP) camera design is proposed to prevent such attacks. DyPP cameras rely on dynamic optical elements, such spatial light modulators, to implement a time-varying PSF, which changes from picture to picture. PSFs are drawn randomly with a learned manifold embedding, trained adversarially to simultaneously meet user-specified targets for privacy, such as face recognition accuracy, and task utility. Empirical evaluations on multiple privacy-preserving vision tasks demonstrate that the DyPP design is significantly more robust to PSF inversion attacks than previous PPCs. Furthermore, the hardware feasibility of the approach is validated by a proof-of-concept camera model.

Keywords: Privacy-preserving Camera

1 Introduction

The joint evolution of cameras and computer vision algorithms enabled the popularization of applications such as crowd monitoring [38], autonomous driving [3,22,30], and smart homes [4,24,67]. However, the increasing deployment of cameras, both at home and in public spaces, raises significant concerns about privacy [6,19,28,29,48,51,53,61]. This literature can be divided into *software-level* and *hardware-level privacy protection*. The former aims to post-process images collected with non-private cameras to guarantee privacy *a posteriori*. Examples include redaction algorithms (*e.g.* face swapping [2, 54] or blur filtering [19]). These approaches maintain the risk that private information could be leaked before post-processing. Hardware-level protection aims to prevent this by guaranteeing that private information is never collected.

The success of end-to-end optical design, know as *deep optics* [1,5,7,8,10,26, 27,32,33,37,42,43,59,60,65], suggests the feasibility of hardware-level privacy.

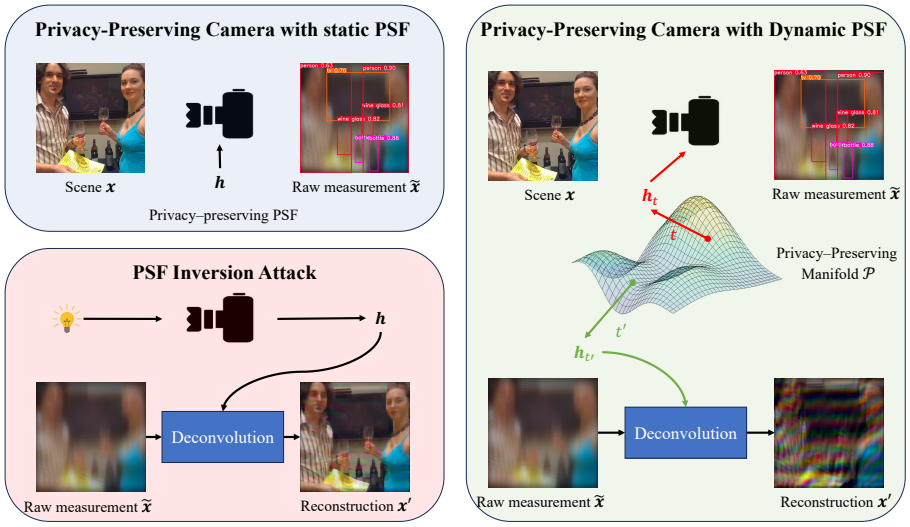

Fig. 1: PPCs that implement a static PSF $\boldsymbol{h}$ enable the solution of tasks like object detection while maintaining privacy (top-left). However, this protection can be easily overridden with PSF inversion attacks. A point light source is used to recover $\boldsymbol{h}$, allowing the recovery of subject identities by simple deconvolution (bottom-left). The DyPP camera uses a dynamic PSF to prevent these attacks (right). A PSF $\boldsymbol{h_t}$ is randomly sampled from the manifold of privacy preserving camera parameters before picture t is taken. This creates a mismatch with a PSF $\boldsymbol{h_{t'}}$, obtained via inversion attacks, preventing the recovery of private information.

In fact, some previous works have addressed the design of privacy-preserving cameras (PPCs) by end-to-end optimization of the camera point spread function (PSF) [28, 29, 61]. This has shown that it is possible to blur images enough to prevent the recognition of faces and other identifying subject traits while still capturing enough information to solve vision tasks such as human pose estimation [29], action recognition [28], or depth estimation [61]. In Figure 1, we provide an example for object detection.

While these works serve as a proof of concept, practical privacy must consider how privacy guarantees can resist a motivated attacker. Although previous works have shown robustness to blind reconstruction techniques, such as Wiener deconvolution [16] or the DeblurGAN [36], this underestimates the risk of reconstruction. As illustrated in Figure 1, an attacker with physical access to the camera can precisely estimate the PSF $\boldsymbol{h}$ by simply measuring the camera response to a point light source. We denote this as a *PSF inversion attack*. Given the PSF, many modern techniques for the solution of inverse problems [64] can produce a reconstruction of reasonably high quality from the blurred images. This is also shown in Figure 1 where the total variation denoising (TVD) algorithm of [56] is used with the camera PSF to reconstruct an image that reveals subject identity. Hence, given access to the camera, it is possible to overcome the privacy guarantees for most scenes.

In this work, we consider the design of *dynamic privacy-preserving* (DyPP) cameras that use a time-varying PSF $\boldsymbol{h}_t$ to prevent PSF inversion attacks. We formalize privacy as an upper bound on the face recognition accuracy of a state-of-the-art recognizer on the camera images. A PSF that meets this privacy guarantee while preserving the scene information necessary to solve vision task $\mathcal{T}$ from its images is denoted privacy-preserving for $\mathcal{T}$. Figure 1 shows an example for the object detection task. The set of all such PSFs is denoted the *privacy manifold* $\mathcal{P}(\mathcal{T})$ of $\mathcal{T}$. As shown in the right inset of the figure, we hypothesize that, for any task $\mathcal{T}$, $\mathcal{P}(\mathcal{T})$ is sufficiently diverse so that an inversion attack based on one PSF $\boldsymbol{h}_t$ is ineffective for images collected with another PSF $\boldsymbol{h}_{t'}$ randomly sampled from $\mathcal{P}(\mathcal{T})$. It follows that, by randomly sampling from $\mathcal{P}(\mathcal{T})$, it is possible to *change* the camera PSF from image to image, thus preventing PSF inversion attacks[3].

An implementation of the DyPP camera is proposed, based on two main contributions. The first is a novel PPC design that relies on an optical device known as a spatial light modulator (SLM) to implement a programmable PSF that changes with each photo. The second is an algorithm that samples PSFs from the *privacy manifold* $\mathcal{P}(\mathcal{T})$, so as to meet the desired privacy guarantees while enabling the solution of task $\mathcal{T}$. This involves the learning of an embedding into the PSF manifold, end-to-end using the camera model, with loss functions that encourage an optimal trade-off between multiple objectives: (1) meeting the target privacy bound for face recognition, (2) maximizing utility for the task, (3) mitigating measurement information that can be used to invert the camera, and (4) maximizing sampled PSF diversity, to prevent PSF inversion attacks.

Experiments show that the DyPP camera can successfully sample a diverse set of PSFs from the privacy manifold to produce privacy-preserving images that enable the solution of several vision tasks, including crowd counting, pose estimation, or object detection. We also report on the construction of a proof-of-concept camera system that validates the practical feasibility of the approach.

2 Related Works

PPCs The preservation of visual privacy has long been studied [49]. Software-level methods post-process images collected by standard cameras to remove sensitive information. Redaction techniques include blur filtering [19, 34, 45, 66], human/object removal [12,13], face swapping [2,54], or visual abstraction [9,18]. Other techniques not explicitly designed for privacy protection, such as style transfer [25], can also be leveraged. The main limitation of this approach is that sensitive information can be leaked before the post-processing. To prevent this risk, there has recently been an interest in *hardware-level* methods that eliminate private information from the raw sensor measurements, while still enabling the solution of vision tasks. This approach relies on special camera or sensor

[3] This assumes that standard protections are used to prevent attackers from hacking into the camera or recovering camera parameters if they do.

designs [53] (*e.g.* extremely low resolution cameras [57, 58, 61], miniature vision sensors [51, 52]). A recent and promising trend is to design the optical elements of a PPC in an end-to-end manner, *i.e.* to jointly optimize the camera and the subsequent vision network to achieve the best trade-off between privacy and task performance. [28, 29] optimize a freeform lens jointly with a utility network for human pose estimation and activity recognition, respectively. [61] learns a privacy-preserving phase mask for passive depth estimation.

Attacks These works have empirically validated robustness to deep blind deconvolution techniques, such as Wiener deconvolution [16] or the DeblurGAN [36]. However, they have not considered PSF inversion attacks, which enable the use of much more powerful deconvolution methods and image reconstruction. This is particularly problematic because the PSF of a linear shift-invariant (LSI) camera can be recovered by simply taking photographs of a point light source. Hence, these attacks are easily within reach of anyone with physical camera access. One possibility to circumvent this risk is to render the camera non-LSI. For example, [57, 58, 61] enforce a extremely low-resolution raw measurement (*e.g.* 16×16) by introducing a downsampling operation at the camera sensor level. However, the extremely low resolution of the raw measurement limits their application in more challenging utility tasks. In this work, we explore the alternative of using a time-varying PSF.

Deep Optics End-to-end optimization of optic designs is now popular in the vision and imaging communities. It has achieved success for applications such as achromatic extended depth of field and super-resolution imaging [59], demosaicing [5], time-of-flight imaging [10, 42, 60], high-dynamic-range imaging [43], microscopy [27, 33], monocular depth estimation [8, 26], and hyperspectral imaging [1, 32, 37]. Recently, it has also been applied to design PPCs [28, 29, 61].

3 DyPP Camera Design Methodology

3.1 Motivation

A PPC aims to capture images that enable the solution of computer vision tasks without revealing private information (*e.g.* subject identity, race). This can be done by end-to-end design of camera optics, with a a differentiable convolutional image formation model [28, 29, 61]. Given scene $\boldsymbol{x}$, the camera measurement is

$$\widetilde{\boldsymbol{x}} = \boldsymbol{h} * \boldsymbol{x} + \boldsymbol{\eta}, \tag{1}$$

where $\boldsymbol{h}$ is the camera PSF computed using Fourier Optics [21], $*$ denotes 2D convolution, and $\boldsymbol{\eta}$ is a vector of i.i.d. white Gaussian noise $\boldsymbol{\eta}_i \sim \mathcal{N}(0, \sigma^2)$.

As illustrated on the left of Figure 1, these approaches can achieve a good trade-off between privacy and computer vision performance. Given original scene projection $\boldsymbol{x}$, the PPC produces a blurry image $\widetilde{\boldsymbol{x}}$ that hides the identity of the scene subjects, while enabling object detection by a modern object detector [41, 62] finetuned on a blurry dataset. However, these works fail to consider the

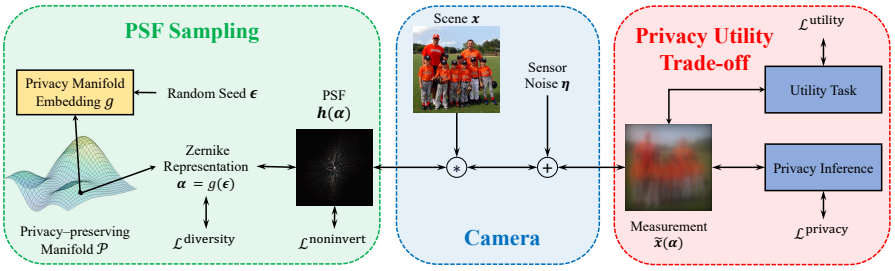

Fig. 2: End-to-end training of the DyPP camera. A manifold embedding g maps a random code ϵ into a vector α of PSF parameters on the privacy manifold $\mathcal{P}$. This produces a PSF $\boldsymbol{h}(\boldsymbol{\alpha})$. Given scene $\boldsymbol{x}$, the camera then produces measurements $\tilde{\boldsymbol{x}}$. The embedding g is trained with task and utility losses that guarantee the balance between privacy (upper bound on face recognition accuracy) and utility (lower bound on target task accuracy) desired for the the privacy manifold $\mathcal{P}$. Double arrows that indicate both forward and backward propagation are performed.

problem that the camera PSF $\boldsymbol{h}$ can be measured by simply introducing an impulse (in practice a point light source) at the camera input. Given the PSF, many deconvolution algorithms can then be used to obtain a reconstruction $\boldsymbol{x'}$ of $\boldsymbol{x}$ from $\widetilde{\boldsymbol{x}}$. We denote this procedure as a *PSF inversion attack*. As illustrated on the right of Figure 1, for the total variation denoising algorithm of [56], the reconstruction usually suffices to identify the subjects in the scene. In this work, we seek a privacy-preserving camera robust against PSF inversion attacks.

3.2 Privacy Manifold

We hypothesize that there are many PSFs that meet the privacy goal, *i.e.* protect subject identity while enabling the solution of a target computer vision task $\mathcal{T}$. We denote the set of such PSFs as the *privacy manifold* of $\mathcal{T}$. Formally, consider the image formation model of (1). Let $\mathcal{X}$ be the set of 2D projections $\boldsymbol{x}$ of the scene, $\widetilde{\mathcal{X}}(\boldsymbol{h})$ the set of images produced by the camera of PSF $\boldsymbol{h}$, $\boldsymbol{\Phi}$ be a face recognizer, and $\rho_{\Phi}(\boldsymbol{h})$ its face recognition accuracy on $\widetilde{\mathcal{X}}(\boldsymbol{h})$.[4] The camera is said to be private at level p if $\rho_{\Phi}(\boldsymbol{h}) < p$. Given a utility network $\boldsymbol{\Psi}$ for task $\mathcal{T}$, the goal is to design the camera that achieves performance $\rho_{\Psi}(\boldsymbol{h}) > \tau$ on $\widetilde{\mathcal{X}}(\boldsymbol{h})$. We denote the set of camera PSFs that meet the two bounds,

$$\mathcal{P}(p,\tau)(\mathcal{T}) = \{\boldsymbol{h} \mid \rho_{\Phi}(\boldsymbol{h}) < p, \rho_{\Psi}(\boldsymbol{h}) > \tau\} \tag{2}$$

as the privacy manifold of parameters (p,τ) for task $\mathcal{T}$. In the remainder of this work, we omit the dependence on the task $\mathcal{T}$ to simplify all equations.

[4] While we focus on face recognition to measure privacy risk, the proposed camera design can be trivially extended to other privacy criteria, such as age, gender, or race classification accuracy.

3.3 Privacy Preserving Camera Approach

The basic idea of the proposed dynamic privacy-preserving (DyPP) camera architecture is to learn how to randomly sample PSFs $\boldsymbol{h} \in \mathcal{P}(p, \tau)$. A different PSF $\boldsymbol{h}_i$ can then be sampled before each image $\widetilde{\boldsymbol{x}}_i$ is acquired. Since this leads to a time-varying PSF $\boldsymbol{h}$, it becomes impossible to recover a generic $\boldsymbol{h}$ with a point light source. However, sampling from the privacy manifold requires the simultaneous satisfaction of the two bounds of (2). This requires a trade-off between eliminating all information that gives away subject identity and preserving all the information needed to the successful completion of the task $\mathcal{T}$. We achieve this goal by relying on a differentiable camera model and end-to-end optimization of the PSF sampling function with respect to these two goals. This is implemented with the architecture of Figure 2, composed by three main blocks: camera, PSF sampling network, and privacy (face recognition) and utility (task) performance trade-off balancing. We next discuss these components.

Differentiable Camera Model Our approach utilizes a spatial light modulator (SLM), an active optical element with controllable phase delay, in the pupil of a conventional imaging lens to produce a shift-invariant imaging system with a dynamic PSF. We parameterize the pupil phase, $\phi_{\boldsymbol{\alpha}}(x, y)$, as

$$\phi_{\boldsymbol{\alpha}}(x, y) = \sum_{i=1}^{d_z} \alpha_i(t) \boldsymbol{Z}_i(x, y), \tag{3}$$

where $\boldsymbol{Z}_i$ is the i-th Zernike polynomial sorted by the Noll's indices [47] and $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^{d_z}$ is the corresponding vector of Zernike coefficients which are used to control the camera PSF at time t. The complex pupil transmittance is given by

$$\boldsymbol{t}_{\boldsymbol{\alpha}}(x, y) = \exp\left[ik\phi_{\boldsymbol{\alpha}}(x, y)\right] \boldsymbol{A}(x, y) \tag{4}$$

where $k = \frac{2\pi}{\lambda}$ is the wavenumber for wavelength λ and $\boldsymbol{A}(x, y)$ models the pupil amplitude transmittance, taking on a value of 1 inside the SLM active area and 0 otherwise. We choose the lens focal length and aperture size such that objects from 1.3m to ∞ are all in-focus. As a result, the wavefront in the pupil from an on-axis point source in the world is planar. The PSF is well approximated by the squared magnitude of the 2D Fourier transform of the pupil transmittance [21]

$$h_{\boldsymbol{\alpha}}(x', y', \lambda) \propto \frac{1}{\lambda^2} \left| \mathcal{F}\left\{\boldsymbol{t}_{\boldsymbol{\alpha}}(x, y)\right\} \Big|_{f_x = \frac{x'}{\lambda f}, f_y = \frac{y'}{\lambda f}} \right|^2. \tag{5}$$

Note that, per Fourier optics theory, the spatial frequency coordinates of the Fourier transform of the pupil, (f_x, f_y), are replaced by $\left(\frac{x'}{\lambda f}, \frac{y'}{\lambda f}\right)$, yielding a function of sensor spatial coordinates (x', y'). Equation 5 is implemented at wavelengths 640 nm, 550 nm and 460 nm using a discrete Fourier transform with appropriate sampling and zero-padding. The resulting PSF is convolved with a training image using (1) to produce a simulated measurement. The measurement is differentiable with respect to the Zernike coefficients, $\boldsymbol{\alpha}$. An illustration of the differentiable camera model is presented in Figure 3.

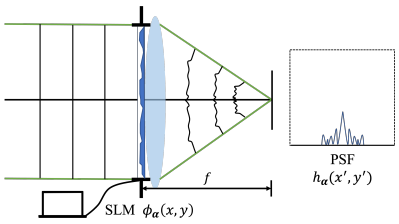

Fig. 3: Image formation model.

PSF Sampling As illustrated on the right of Figure 2, the privacy manifold is modeled with a non-linear embedding

$$g : \Delta \rightarrow \mathcal{A} \tag{6}$$

from an interval $\Delta = [-1, 1]^{d_z}$ of latent codes to the space $\mathcal{A}$ of Zernike coefficient vectors $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{d_z})^T$. This embedding is trained with loss functions that encourage camera privacy ($\mathcal{L}^{\text{noninvert}}$) and PSF diversity ($\mathcal{L}^{\text{diversity}}$). After training, a PSF $\boldsymbol{h}$ in the privacy manifold $\mathcal{P}$ is obtained by sampling a random code $\boldsymbol{\epsilon} \sim \mathcal{U}(\Delta)$ from a multivariate uniform distribution over Δ, obtaining a Zernike coefficients vector with

$$\boldsymbol{\alpha} = g(\boldsymbol{\epsilon}) \tag{7}$$

and synthesizing the PSF $\boldsymbol{h}$ with (3)-(5). We next discuss the losses in detail.

Noninvertibility Loss: This loss aims to increase the difficulty of reconstructing the scene $\boldsymbol{x}$ from the camera measurements $\widetilde{\boldsymbol{x}}$. It leverages the fact that, given a linear observation model $\widetilde{\boldsymbol{x}} = \boldsymbol{H}\boldsymbol{x} + \eta$, the mean squared error (MSE) ϵ_{MSE} of the MMSE estimator of $\boldsymbol{x}$ given $\widetilde{\boldsymbol{x}}$ is [23]

$$\epsilon_{\text{MSE}} \propto \sigma^2 \text{tr}\left((\boldsymbol{H}\boldsymbol{H}^*)^{-1}\right). \tag{8}$$

Since the camera model is a 2D convolution, $\boldsymbol{H}$ is a doubly block circulant matrix diagonalizable by the 2D discrete Fourier transform (DFT) matrix $\mathcal{F}$, according to $\boldsymbol{H} = \mathcal{F}^*\boldsymbol{\Lambda}\mathcal{F}$. Here, $\boldsymbol{\Lambda} = \text{diag}(\mathcal{F}\{h\})$ is a diagonal matrix containing the 2D DFT of $\boldsymbol{h}$. Substituting into (8) it can be shown that the MSE of the MMSE estimator is given by

$$\epsilon_{\text{MSE}} = \sigma^2 \sum_i \frac{1}{|\boldsymbol{\Lambda}_{ii}|^2}. \tag{9}$$

Hence, we use a loss which maximizes the reconstruction MSE by minimizing

$$\mathcal{L}^{\text{noninvert}} = \sum_i -\frac{1}{|\boldsymbol{\Lambda}_{ii}|^2 + \varepsilon} \tag{10}$$

where $\varepsilon > 0$ is a small constant, added for numerical stability. Note that this loss encourages the minimization of the coefficients $\{\boldsymbol{\Lambda}_{ii}\}_i$ of the Fourier transform of $\boldsymbol{h}$. Therefore, the noninvertibility loss can be seen as a regularizer that

encourages smooth PSFs $\boldsymbol{h}$, leading to blurry images and increased difficulty of image reconstruction.

Diversity Loss: This loss encourages the privacy manifold $\mathcal{P}$ to contain a large diversity of PSFs, so as to induce more variability from PSF to PSF, increase the difficulty of measuring the PSF, and thus prevent PSF inversion attacks. While the Shannon entropy of the random vector $\boldsymbol{\alpha}$ of Zernike coefficients is a sensible diversity measure, it is difficult to derive a differentiable and computationally efficient estimator of this entropy [50]. To circumvent this, we instead maximize the upper bound [11] $H(\boldsymbol{\alpha}) \leq \frac{1}{2}\log\det\mathrm{cov}(\boldsymbol{\alpha}) + \frac{d_z}{2}\log(2\pi e)$. This leads to the *diversity loss*

$$\mathcal{L}^{\text{diversity}} = -\frac{1}{2}\log\det\widehat{\mathrm{cov}}(\boldsymbol{\alpha}) \tag{11}$$

where $\widehat{\mathrm{cov}}(\boldsymbol{\alpha})$ is the sample covariance of a set of N^{cov} randomly sampled Zernike representations $\{\boldsymbol{\alpha}_i = g(\boldsymbol{\epsilon}_i)\}_{i=1}^{N^{\text{cov}}}$.

Privacy-utility Trade-off So far, we have discussed how to sample a set of diverse and non-invertible PSFs. The final module of Figure 2 aims to ensure that these PSFs satisfy the privacy manifold constraints of (2), in terms of face recognition accuracy upper bound p and task performance accuracy τ.

Privacy Loss: The closed-set face recognition performance of a state-of-the-art recognizer [40] is used as a proxy privacy criterion for camera design. The face recognizer is a multiclass predictor $\Phi : \mathcal{X} \to \mathbb{R}^C$. Given a face image $\boldsymbol{x}$ it returns subject identity $y^* = \arg\max_j \Phi_j(x)$.

Given a face recognition dataset $\mathcal{D}^{\text{P}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N^{\text{P}}}$ consisting of N^{P} face images from C identities, the images are passed through the camera of PSF coefficients $\boldsymbol{\alpha}$ to obtain measurements $\{\widetilde{\boldsymbol{x}}_i(\boldsymbol{\alpha})\}_{i=1}^{N^{\text{P}}}$. Using the definition of classification margin

$$\mathcal{M}(y, \Phi(\boldsymbol{x})) = \Phi_y(\boldsymbol{x}) - \max_{j \neq y} \Phi_j(\boldsymbol{x}), \tag{12}$$

the recognition accuracy of Φ can be estimated by

$$\widehat{\rho}_\Phi(\boldsymbol{\alpha}) = 1 - \frac{1}{N^{\text{P}}} \sum_{(\boldsymbol{x}_i, y_i) \sim \mathcal{D}^{\text{P}}} \ell_{0/1}\left(\mathcal{M}(y_i, \Phi(\widetilde{\boldsymbol{x}}_i(\boldsymbol{\alpha})))\right), \tag{13}$$

where $\ell_{0/1}(z) = (1 - \mathrm{sign}(z))/2$ is the $0/1$ loss function. This loss identifies misclassifications, *i.e.* instances with negative margin but is non-differentiable and intractable for gradient-based optimization. We rely on the approximation by the sigmoid loss $\ell_{sig}(z) = (1 + \exp(\eta z))^{-1}$ where $\eta > 0$ is a smoothness parameter.

To encourage the PSF $\boldsymbol{h}(\boldsymbol{\alpha})$ to meet the privacy requirement of (2), we rely on *privacy loss*

$$\mathcal{L}^{\text{privacy}} = \max(\widehat{\rho}_\Phi(\boldsymbol{\alpha}) - p, 0). \tag{14}$$

This is minimized when $\widehat{\rho}_\Phi(\boldsymbol{\alpha}) \leq p$, *i.e.* the privacy requirement is met.

Utility Loss: The performance of a state-of-the-art algorithm/network Ψ for the task $\mathcal{T}$ is used as proxy utility criterion for the DyPP camera. Given a dataset

$\mathcal{D}^{\mathrm{U}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N^{\mathrm{U}}}$ of task $\mathcal{T}$, the performance of $\boldsymbol{\Psi}$ can be estimated by

$$\widehat{\rho}_{\boldsymbol{\Psi}}(\boldsymbol{\alpha}) = \frac{1}{N^{\mathrm{U}}} \sum_{(\boldsymbol{x}_i, y_i) \sim \mathcal{D}^{\mathrm{U}}} -u^{\boldsymbol{\Psi}}(\widetilde{\boldsymbol{x}}_i, y_i) \tag{15}$$

where $u^{\boldsymbol{\Psi}}$ is a surrogate loss function for the task. To encourage the PSF $\boldsymbol{h}(\boldsymbol{\alpha})$ to meet the task utility requirement of (2), we rely on *utility loss*

$$\mathcal{L}^{\mathrm{utility}} = \max\left(\tau - \widehat{\rho}_{\boldsymbol{\Psi}}(\boldsymbol{\alpha}), 0\right). \tag{16}$$

This is minimized when $\widehat{\rho}_{\boldsymbol{\Psi}}(\boldsymbol{\alpha}) \geq \tau$, *i.e.* the task performance requirement is met. It is also possible to learn a privacy manifold compatible with multiple utility tasks (*e.g.* a multi-purpose PPC) by simply summing (16) across tasks.

By minimizing $\mathcal{L}^{\mathrm{privacy}}$ and $\mathcal{L}^{\mathrm{utility}}$ jointly and adversarially, the generated PSFs are expected to preserve as much information utilizable for the vision tasks as possible while meeting the privacy requirement. Note that by setting $\tau = \infty$ it is possible to simply optimize the task utility while meeting the privacy guarantee. This encourages the best possible task performance under the privacy guarantee. We use this setting in all our experiments. Overall, the privacy manifold embedding g is trained by jointly optimizing the combination of all losses in (10), (11), (14), and (16), *i.e.* using the loss

$$\mathcal{L}^{\mathrm{overall}} = \lambda_1 \mathcal{L}^{\mathrm{noninvert}} + \lambda_2 \mathcal{L}^{\mathrm{diversity}} + \lambda_3 \mathcal{L}^{\mathrm{utility}} + \lambda_4 \mathcal{L}^{\mathrm{privacy}}, \tag{17}$$

where the coefficients $\{\lambda_i\}_{i=1}^{4}$ are hyperparameters.

Training A Zernike vector $\boldsymbol{\alpha}$ is randomly sampled per training iteration, using (7), for the purposes of computing $\mathcal{L}^{\mathrm{noninvert}}$, $\mathcal{L}^{\mathrm{privacy}}$, and $\mathcal{L}^{\mathrm{utility}}$. To compute $\mathcal{L}^{\mathrm{diversity}}$, we additionally sample a set of Zernike vectors $\{\boldsymbol{\alpha}_i = g(\boldsymbol{\epsilon}_i)\}_{i=1}^{N^{\mathrm{cov}}}$ per iteration to derive sample covariance $\widehat{\mathrm{cov}}(\boldsymbol{\alpha})$. We found in practice that when the utility network $\boldsymbol{\Psi}$ is trained purely on clear images, the utility performance estimate of (15) is unrealistically low. To avoid this, we unfreeze $\boldsymbol{\Psi}$ during training and adapt its parameters of by back-propagating $\mathcal{L}^{\mathrm{utility}}$ through the network.

4 Experiments

In this section, we empirically evaluate the ability of the DyPP camera design to meet the privacy and utility bounds of (2) and its robustness to PSF inversion attacks. In addition, we discuss a proof-of-concept physical camera model built to demonstrate the hardware feasibility of the approach.

4.1 Experimental Set-up

Optics Simulation We simulate a camera sensor with a pixel size of 1 µm, an f-number of 1.8 and a resolution size of 640×640 pixels. Following the protocol of [29, 59], the number of Zernike coefficients is set to $d_z = 350$.

Table 1: Utility performance evaluation networks and datasets.

Task	Network Ψ	Dataset $\mathcal{D}^{\text{utility}}$
Crowd Count	MAN [38]	ShanghaiTech Part B [68] Beijing BRT [15]
Pose Est.	YOLO-Pose [41]	COCO Keypoints 2017 [39]
Obj. Det.	YOLOv5 [62]	COCO Detection 2017 [39]

Table 2: Performance on different privacy inference tasks.

	closed-set face recognition	face verification	
	PubFig [35]	LFW [31]	AgeDB-30 [44]
	Accuracy	AUROC	
No Privacy	0.96±0.01	1.00±0.00	0.99±0.01
Low-resolution [57, 58]	**0.07±0.01**	0.75±0.02	**0.62±0.02**
Defocus Lens [51]	0.14±0.01	0.76±0.03	0.65±0.02
PP-HPE Lens [29]	0.09±0.01	0.73±0.03	0.63±0.03
DyPP	0.09±0.02	**0.72±0.02**	0.63±0.03

Privacy Criterion The face recognition privacy criterion is evaluated on a subset of the PubFig [35] dataset of public figure faces. This includes 175 identities, each with 25 training images and 5 test images[5]. All the training images are used as $\mathcal{D}^{\text{P}}$ in (14). An off-the-shelf ArcFace [14] face recognizer with IResNet-100 [17] backbone is used as the privacy inference model in $\mathcal{L}^{\text{privacy}}$. We set the manifold privacy bound of (2) to $p = 0.1$.

Utility Criteria Utility performance is evaluated on three tasks: crowd counting, pose estimation, and object detection. The datasets and networks used for evaluation are summarized in Table 1. For all networks, we use the official public implementations. We set the manifold utility bound of (2) $\tau = \infty$ for all three tasks. As discussed above, this encourages the maximization of task utility under the privacy constraint set by p.

Reconstruction Attacks The following attack strategies are considered.

Blind: attacker without access to camera hardware, reconstructs images by blind deconvolution via Deep Image Prior [63].

Deep Learning: attacker without access to camera hardware but access to dataset of image pairs, each including an image collected with the camera and a clear image of the scene. Attacker trains an encoder-decoder deblurring U-net [55] with skip connections, following [20, 46].

PSF inversion: attacker with access to camera, performs PSF inversion attack and uses recovered PSF in the total variation denoising (TVD) algorithm [56].

White- vs Black-Box: For PSF inversion attacks, the PSF $\boldsymbol{h}$ of a static camera can be easily measured using a point light source. This is denoted as a *white-box* attack. The time-varying PSF $\boldsymbol{h}_t$ of the DyPP camera prevents white-box attacks. The attacker can still use a point light source to measure a set $\mathcal{H} = \{\boldsymbol{h}_i\}_i$ of camera PSFs. Given a private image, the attack can then be performed with all PSFs in $\mathcal{H}$ and the best reconstruction chosen, *e.g.* by visual inspection.

[5] We ignored 25 identities without enough face image samples, due to invalid URLs.

Table 3: Pose estimation performance on the COCO validation set.

	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Low-resolution [57, 58]	10.3	26.2	6.5	5.9	16.9	14.4
Defocus Lens [51]	30.8	58.9	34.1	15.9	51.6	34.1
PP-HPE Lens [29]	**42.8**	**69.8**	**44.2**	**29.8**	**59.9**	**50.1**
DyPP	42.2	69.3	43.6	29.5	59.1	49.7

Table 4: Object detection performance on the COCO validation set.

	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Low-resolution [57, 58]	8.2	21.4	6.9	5.7	14.0	13.8
Defocus Lens [51]	24.9	40.6	26.7	29.7	41.3	22.0
PP-HPE Lens [29]	33.9	51.0	35.9	37.0	50.2	28.5
DyPP	**36.0**	**53.7**	**38.4**	**39.2**	**51.3**	**30.5**

Table 5: Crowd counting performance on ShanghaiTech B and Beijing BRT test sets.

	ShanghaiTech B [68] MAE/MSE	Beijing BRT [15] MAE/MSE
Low-resolution [57, 58]	41.4/60.5	10.65/13.90
Defocus Lens [51]	25.3/42.4	3.28/4.50
PP-HPE Lens [29]	21.3/33.7	2.93/4.12
DyPP	**19.3**/**28.7**	**2.35**/**3.43**

This is denoted as a *black-box* attack. White-box attacks are simulated by using the true PSF $\boldsymbol{h}$ in the TVD algorithm. For black-box attacks this is replaced by a set of PSFs randomly sampled from the the privacy manifold. The best reconstruction is chosen by measuring PSNR/SSIM of the reconstruction.

Baselines The proposed DyPP camera is compared to two previous LSI PPCs of static PSF: Defocus lens [51] and PP-HPE lens [29]. For PSF inversion attacks, the robustness of the baselines to white-box attacks is compared to that of the DyPP camera to black-box attacks. In addition, we consider a camera with extremely low resolution (32×32) [57, 58] as a non-LSI privacy camera baseline.

Training Details The manifold embedding network g used to implement the DyPP camera is a 4-layer MLP with leaky ReLU activation and 512 nodes for each hidden layer. The network is trained in two stages. In the first stage, we train a generic manifold embedding g over the three tasks, using an image batch size of 128 for $\mathcal{L}^{\text{privacy}}$ and a batch of size 8/8/1 (object detection/pose estimation/crowd counting) for $\mathcal{L}^{\text{utility}}$. Training is performed by a stochastic gradient descent optimizer with learning rate of 1e-4 for 10,000 iterations, which takes roughly 3 days on an NVIDIA-A100-80GB GPU. In the second stage, we further finetune the utility networks with blurry images generated by the trained g. For black-box PSF inversion attacks on DyPP, we use a set of $|\mathcal{H}| = 64$ randomly sampled PSFs by default.

4.2 PPC Performance

Tables 2 - 5 summarize the trade-off between privacy and task utility achieved by all PPCs considered in this work.

Fig. 4: Qualitative results on different utility computer vision tasks. From left to right: original image, DyPP image, task outputs. From top to bottom: object detection, pose estimation, crowd counting.

Table 6: Robustness to reconstruction attacks. Higher reconstruction quality in terms of SSIM/PSNR indicates lower robustness.

	Raw		Blind [63]		Deep Learning [55]		white-box TVD [56]		Black-box TVD [56]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Defocus Lens [51]	16.6	0.601	17.3	0.645	16.2	0.626	22.1	0.788	-	-
PP-HPE Lens [29]	14.7	0.558	16.9	0.623	15.9	0.601	21.8	0.756	-	-
DyPP	14.8	0.571	16.1	0.589	15.6	0.584	20.3	0.745	17.3	0.638

Face Recognition Accuracy Table 2 summarizes privacy performance in terms of face recognition accuracy. This is 96% on the clear images of the PubFig test set, but decreases to less than 15% on images from all cameras, showing that they are effective at guaranteeing privacy. To investigate how privacy generalizes to other tasks, we measured performance on two face-verification tasks using the LFW [31] and AgeDB-30 [44] datasets. Note that the cameras were not retrained on these dataset. Table 2 shows that the results were qualitatively similar to those of PubFig. All cameras significantly degraded the nearly perfect AUROC of the original images. Among the different cameras, Low-resolution sensors are the most private on PubFig, but DyPP generalized better to LFW, and the two cameras have similar performance on AgeDB-30. Note that privacy does not matter in isolation, as the images collected by the camera must preserve enough information to allow the solution of the vision task.

Utility Performance Tables 3-5 summarize the utility performance of all PPC designs on the three tasks considered. In all tasks, the performance of both the Low-resolution sensor and Defocus Lens was poor. DyPP and PP-HPE had comparable pose estimation performance on COCO, but DyPP outperformed all other cameras for object detection and crowd counting.

Utility-privacy Trade-off Altogether, it can be concluded that Low-resolution sensors achieve privacy by eliminating too much information, degrading the performance of the PPC for the vision tasks considered. Due to this weak perfor-

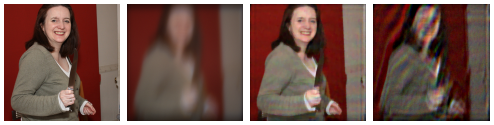

Fig. 5: From left to right: clear image, DyPP raw measurement, white-box TVD reconstruction, and black-box TVD reconstruction.

Table 7: Performance on different privacy inference tasks.

	PubFig [35] Accuracy	LFW [31] AUROC	AgeDB-30 [44] AUROC
Defocus Lens [51]	0.14±0.01	0.76±0.03	0.65±0.02
+ White-box TVD	0.72±0.02	0.86±0.02	0.79±0.02
PP-HPE Lens [29]	0.09±0.01	0.73±0.03	0.63±0.03
+ White-box TVD	0.65±0.03	0.84±0.02	0.78±0.03
DyPP	0.09±0.02	0.72±0.02	0.63±0.03
+White-box TVD	0.54±0.03	0.81±0.03	0.75±0.03
+ Black-box TVD	0.14±0.04	0.75±0.03	0.68±0.02

mance we do not consider these sensors in the remaining experiments. Among the LSI cameras, PP-HPE has a better trade-off than Defocus. DyPP outperforms PP-HPE on crowd counting and object detection, but slightly underperforms on pose estimation. This is unsurprising given that the later is specifically optimized for pose estimation [29].

Qualitative Results Figure 4 shows examples of original images, images captured with the DyPP camera and the results of object detection, pose estimation, and crowd counting on the latter. It is interesting that the computer vision networks perform quite well on images that are degraded to the point of being nearly unintelligible for a human.

Robustness to PSF Inversion Attacks We next consider the robustness of the different LSI cameras to PSF inversion attacks. We report white-box attack performance for all PPCs. Note that the black-box attack is the only practical attacks for DyPP, while the white-box attack is infeasible for DyPP, whose point is to prevent them. We account for this by shading white-box results in the table.

Reconstruction Quality Table 6 summarizes the robustness of the LSI cameras in terms of image reconstruction metrics like SSIM and PSNR between original and reconstructed image. In both cases, lower values denote weaker reconstruction and better privacy protection. "Raw" denotes the absence of attack, showing that all cameras have comparable performance in this setting. The remainder of the table summarizes performance under the different attack strategies. Two conclusions can be taken. First, white-box PSF inversion attacks are much more effective than those previously studied in the literature. For all cameras, blind and deep learning attacks barely increase the raw PSNR/SSIM. This is unlike white-box PSF inversion, which increases PSNR from ~15 to ~20 and SSIM from ~0.55 to ~0.75. Second, black-box inversion attacks are much less effective than their white-box counterparts, and again unable to significantly increase on the PSNR/SSIM of the raw measurement. Figure 5 visualizes an example of images recovered from the DyPP sensor measurement by white-box and black-box PSF inversion attacks. These results show that, by preventing white-box attacks, DyPP is much more privacy-preserving than all baselines.

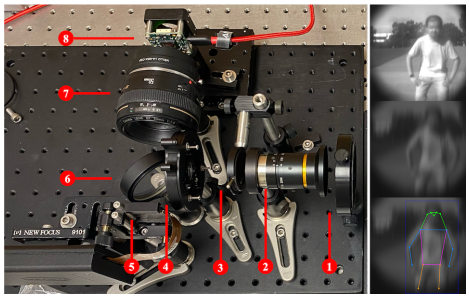

Fig. 6: Camera setup and sample images. Left: Real-world privacy camera implementation: 1. linear polarizer, 2. Kowa 5 mm, f/1.8 lens, 3. Iris as a field stop, 4. 50mm achromatic lens, 5. HOLOEYE PLUTO SLM, 6. beam splitter, 7. 50 mm Canon camera lens, 8. Sony IMX178 board level sensor. Right: (from top to bottom) The ground truth image, privacy-preserving measurement, Privacy-preserving pose estimation.

Face Recognition Accuracy Table 7 extends the face recognition performance characterization of Table 2 to the attack setting. Again, white-box attacks are much more effective than black-box ones. While the PubFig recognition accuracy increases from 9% to 54% for the former, it remains at 14% for the latter. Although this is less private than the raw measurements, the DyPP camera is significantly more robust to inversion attacks than the other LSI PPCs. Similar results hold for LFW [31] and AgeDB-30 [44], where black-box attacks are much less effective than white-box ones.

4.3 Hardware Proof-of-concept

We validate the DyPP concept in a benchtop hardware prototype, shown 6. In principle, any hardware that allows programmable pupil phase could be used. Ours uses a *liquid crystal on silicon* (LCOS) phase-only SLM (HOLOEYE PLUTO), which operates in a reflective geometry, to implement our dynamic pupil phase functions. The incident light is filtered to be quasi-monochromatic and polarized which his required by our particular SLM model. A 5 mm focal, f/1.8 main lens forms a clean, un-blurred image at its back focal plane. We use a relay system comprising a lens and beam splitter to project the main lens' pupil onto the SLM which then imparts our desired phase. Another 50 mm lens (element 7) collects the phase-modulated light that reflects from the SLM, forming an image on the sensor that is blurred by our prescribed PSF.

5 Conclusion

In this work, we propose a new design of PPC whose PSF is randomly sampled from a privacy-preserving manifold in the parameter space. Due to the time-varying nature of its PSF, this PPC design is significantly more robust to image reconstruction attack, compared to prior PPCs with static PSF.

Acknowledgements We thank Carlos Hinojosa for sharing the PSF of PP-HPE. This work was partially funded by NSF award IIS-2303153, a gift from Qualcomm, and NVIDIA GPU donations. We also acknowledge and thank the use of the Nautilus platform for some of the experiments discussed above.

References

1. Baek, S.H., Ikoma, H., Jeon, D.S., Li, Y., Heidrich, W., Wetzstein, G., Kim, M.H.: Single-shot hyperspectral-depth imaging with learned diffractive optics. In: ICCV (2021)
2. Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., Nayar, S.K.: Face Swapping: Automatically Replacing Faces in Photographs. ACM Transactions on Graphics (ToG) (2008)
3. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., Zieba, K.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016)
4. Bouchabou, D., Nguyen, S.M., Lohr, C., LeDuc, B., Kanellos, I.: A survey of human activity recognition in smart homes based on iot sensors algorithms: Taxonomies, challenges, and opportunities with deep learning. Sensors **21**(18), 6037 (2021)
5. Chakrabarti, A.: Learning sensor multiplexing design through back-propagation. In: NeurIPS (2016)
6. Chan, A.B., Liang, Z.S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: Counting people without people models or tracking. In: CVPR. pp. 1–7. IEEE (2008)
7. Chang, J., Sitzmann, V., Dun, X., Heidrich, W., Wetzstein, G.: Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. Scientific reports **8**(1), 12324 (2018)
8. Chang, J., Wetzstein, G.: Deep optics for monocular depth estimation and 3d object detection. In: ICCV (2019)
9. Chinomi, K., Nitta, N., Ito, Y., Babaguchi, N.: Prisurv: Privacy protected video surveillance system using adaptive visual abstraction. In: Advances in Multimedia Modeling: 14th International Multimedia Modeling Conference, MMM 2008, Kyoto, Japan, January 9-11, 2008. Proceedings 14. pp. 144–154. Springer (2008)
10. Chugunov, I., Baek, S.H., Fu, Q., Heidrich, W., Heide, F.: Mask-tof: Learning microlens masks for flying pixel correction in time-of-flight imaging. In: CVPR (2021)
11. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons (1999)
12. Criminisi, A., Perez, P., Toyama, K.: Object removal by exemplar-based inpainting. In: CVPR (2003)
13. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on image processing (TIP) **13**(9), 1200–1212 (2004)
14. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
15. Ding, X., Lin, Z., He, F., Wang, Y., Huang, Y.: A deeply-recursive convolutional network for crowd counting. In: ICASSP (2018)
16. Dong, J., Roth, S., Schiele, B.: Deep wiener deconvolution: Wiener meets deep learning for image deblurring. In: NeurIPS (2020)
17. Duta, I.C., Liu, L., Zhu, F., Shao, L.: Improved residual networks for image and video recognition. In: ICPR. IEEE (2021)
18. Fan, J., Luo, H., Hacid, M.S., Bertino, E.: A novel approach for privacy-preserving video sharing. In: ACM international conference on Information and knowledge management (CIKM) (2005)

19. Frome, A., Cheung, G., Abdulkader, A., Zennaro, M., Wu, B., Bissacco, A., Adam, H., Neven, H., Vincent, L.: Large-scale privacy protection in google street view. In: ICCV (2009)
20. Gao, H., Tao, X., Shen, X., Jia, J.: Dynamic scene deblurring with parameter selective sharing and nested skip connections. In: CVPR (2019)
21. Goodman, J.W.: Introduction to Fourier Optics. W. H. Freeman, 4th edn. (2017)
22. Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: A survey of deep learning techniques for autonomous driving. Journal of Field Robotics **37**(3), 362–386 (2020)
23. Harwit, M., Sloane, N.J.A.: Hadamard transform optics (1979)
24. Hasnain R., M., S., R., P., M., G., S.: Smart home automation using computer vision and segmented image processing. In: 2019 International Conference on Communication and Signal Processing (ICCSP) (2019)
25. Hassan; Rakibul Hasan; Patrick Shaffer; David Crandall; Apu Kapadia, E.T.: Cartooning for enhanced privacy in lifelogging and streaming videos. In: CVPR Workshops (2017)
26. He, L., Wang, G., Hu, Z.: Learning depth from single images with deep neural network embedding focal length. IEEE Transactions on Image Processing (TIP) **27**(9), 4676–4689 (2018)
27. Hershko, E., Weiss, L.E., Michaeli, T., Shechtman, Y.: Multicolor localization microscopy and point-spread-function engineering by deep learning. Optics express **27**(5), 6158–6183 (2019)
28. Hinojosa, C., Marquez, M., Arguello, H., Adeli, E., Fei-Fei, L., Niebles, J.C.: Privhar: Recognizing human actions from privacy-preserving lens. In: ECCV (2022)
29. Hinojosa, C., Niebles, J.C., Arguello, H.: Learning privacy-preserving optics for human pose estimation. In: ICCV (2021)
30. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., Lu, L., Jia, X., Liu, Q., Dai, J., Qiao, Y., Li, H.: Planning-oriented autonomous driving. In: CVPR (2023)
31. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
32. Jeon, D.S., Baek, S.H., Yi, S., Fu, Q., Dun, X., Heidrich, W., Kim, M.H.: Compact snapshot hyperspectral imaging with diffracted rotation. ACM Transactions on Graphics (ToG) **38**(4) (jul 2019)
33. Kellman, M., Bostan, E., Chen, M., Waller, L.: Data-driven design for fourier ptychographic microscopy. In: IEEE International Conference on Computational Photography (ICCP). pp. 1–8. IEEE (2019)
34. Kitahara, I., Kogure, K., Hagita, N.: Stealth vision for protecting privacy. In: ICPR. vol. 4, pp. 404–407. IEEE (2004)
35. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV (2009)
36. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: Deblurgan: Blind motion deblurring using conditional adversarial networks. In: CVPR (2018)
37. Li, L., Wang, L., Song, W., Zhang, L., Xiong, Z., Huang, H.: Quantization-aware deep optics for diffractive snapshot hyperspectral imaging. In: CVPR (2022)
38. Lin, H., Ma, Z., Ji, R., Wang, Y., Hong, X.: Boosting crowd counting via multifaceted attention. In: CVPR (2022)
39. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)

40. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: CVPR (2017)
41. Maji, D., Nagori, S., Mathew, M., Poddar, D.: Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In: CVPRW (2022)
42. Marco, J., Hernandez, Q., Munoz, A., Dong, Y., Jarabo, A., Kim, M.H., Tong, X., Gutierrez, D.: Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. ACM Transactions on Graphics (ToG) **36**(6), 1–12 (2017)
43. Metzler, C.A., Ikoma, H., Peng, Y., Wetzstein, G.: Deep optics for single-shot high-dynamic-range imaging. In: CVPR (2020)
44. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: CVPRW (2017)
45. Neustaedter, C., Greenberg, S., Boyle, M.: Blur filtration fails to preserve privacy for home-based video conferencing. ACM Trans. Comput.-Hum. Interact. **13**(1), 1–36 (2006)
46. Nimisha, T.M., Kumar Singh, A., Rajagopalan, A.N.: Blur-invariant deep learning for blind-deblurring. In: ICCV (2017)
47. Noll, R.J.: Zernike polynomials and atmospheric turbulence. Journal of the Optical Society of America **66**(3), 207–211 (1976)
48. Orekondy, T., Schiele, B., Fritz, M.: Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In: ICCV (2017)
49. Padilla-López, J.R., Chaaraoui, A.A., Flórez-Revuelta, F.: Visual privacy protection methods: A survey. Expert Systems with Applications **42**(9), 4177–4195 (2015)
50. Pichler, G., Colombo, P.J.A., Boudiaf, M., Koliander, G., Piantanida, P.: A differential entropy estimator for training neural networks. In: ICML (2022)
51. Pittaluga, F., Koppal, S.J.: Privacy preserving optics for miniature vision sensors. In: CVPR (2015)
52. Pittaluga, F., Koppal, S.J.: Pre-capture privacy for small vision sensors. IEEE transactions on pattern analysis and machine intelligence (TPAMI) **39**(11), 2215–2226 (2016)
53. Pittaluga, F., Zivkovic, A., Koppal, S.J.: Sensor-level privacy for thermal cameras. In: 2016 IEEE International Conference on Computational Photography (ICCP). pp. 1–12. IEEE (2016)
54. Ren, Z., Lee, Y.J., Ryoo, M.S.: Learning to anonymize faces for privacy preserving action detection. In: ECCV (2018)
55. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. Springer (2015)
56. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena **60**(1-4), 259–268 (1992)
57. Ryoo, M., Kim, K., Yang, H.: Extreme low resolution activity recognition with multi-siamese embedding learning. In: AAAI (2018)
58. Ryoo, M., Rothrock, B., Fleming, C., Yang, H.J.: Privacy-preserving human activity recognition from extreme low resolution. In: AAAI (2017)
59. Sitzmann, V., Diamond, S., Peng, Y., Dun, X., Boyd, S., Heidrich, W., Heide, F., Wetzstein, G.: End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. ACM Transactions on Graphics (TOG) **37**(4), 1–13 (2018)
60. Su, S., Heide, F., Wetzstein, G., Heidrich, W.: Deep end-to-end time-of-flight imaging. In: CVPR (2018)
61. Tasneem, Z., Milione, G., Tsai, Y.H., Yu, X., Veeraraghavan, A., Chandraker, M., Pittaluga, F.: Learning phase mask for privacy-preserving passive depth estimation. In: ECCV (2022)

62. Ultralytics: YOLOv5: A state-of-the-art real-time object detection system. https://docs.ultralytics.com (2021)
63. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: CVPR (2018)
64. Vogel, C.R.: Computational methods for inverse problems. SIAM (2002)
65. Wetzstein, G., Ikoma, H., Metzler, C., Peng, Y.: Deep optics: Learning cameras and optical computing systems. In: 2020 54th Asilomar Conference on Signals, Systems, and Computers. pp. 1313–1315. IEEE (2020)
66. Wu, Z., Wang, H., Wang, Z., Jin, H., Wang, Z.: Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **44**(4), 2126–2139 (2020)
67. Yu, J., de Antonio, A., Villalba-Mora, E.: Deep learning (cnn, rnn) applications for smart homes: a systematic review. Computers **11**(2), 26 (2022)
68. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: CVPR (2016)