Learning Privacy-preserving Optics for Human Pose Estimation

Carlos Hinojosa[1], Juan Carlos Niebles[2], Henry Arguello[1]
[1]Universidad Industrial de Santander [2]Stanford University

https://carloshinojosa.me/project/privacy-hpe/

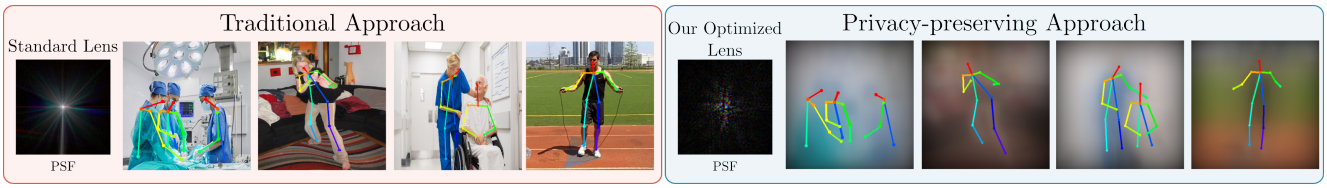

Figure 1: Standard cameras acquire visual details from the scene that could lead to privacy issues. In this work, we propose to learn privacy-preserving optics to perform human pose estimation (HPE). Our optimized lens incorporates several optical aberrations that degrade the image to hide private visual details while it still captures enough visual information to perform human pose estimation.

Abstract

The widespread use of always-connected digital cameras in our everyday life has led to increasing concerns about the users' privacy and security. How to develop privacy-preserving computer vision systems? In particular, we want to prevent the camera from obtaining detailed visual data that may contain private information. However, we also want the camera to capture useful information to perform computer vision tasks. Inspired by the trend of jointly designing optics and algorithms, we tackle the problem of privacy-preserving human pose estimation by optimizing an optical encoder (hardware-level protection) with a software decoder (convolutional neural network) in an end-to-end framework. We introduce a visual privacy protection layer in our optical encoder that, parametrized appropriately, enables the optimization of the camera lens's point spread function (PSF). We validate our approach with extensive simulations and a prototype camera. We show that our privacy-preserving deep optics approach successfully degrades or inhibits private attributes while maintaining important features to perform human pose estimation.

1. Introduction

Cameras are ubiquitous and pervasive today: we have them in our smartphones, cars, homes, and cities. The tremendous amount of data collected from these devices enables a myriad of applications using computer vision-based technologies. We encounter such technologies in our daily life. At hospitals, visual sensors have given rise to ambient intelligence: physical spaces that are sensitive and responsive to the presence of humans. In this scenario, visual systems enable more efficient clinical workflows and improved patient safety in intensive care units and operating rooms [15]. In the context of gaming, camera devices use action and gesture recognition to create an interactive game experience [44, 50]. However, with all these cameras collecting images in an always-connected digital world, a big challenge has been raised: how to develop privacy-preserving computer vision systems? Specifically, we want to prevent the camera system from obtaining detailed visual data containing private information (such as faces), desirably at the hardware level. Simultaneously, we want the system to capture useful information that enables understanding surrounding objects and ongoing events.

For decades, cameras have been engineered to imitate the human vision system. Once the optical system is fixed, we use the cameras to acquire multiple high-fidelity images. Then we tune computer vision algorithms to optimize their accuracy at specific tasks. Most computer vision applications, even privacy-preserving approaches, rely on such a traditional digital imaging system. For example, one can detect privacy-sensitive everyday situations and enable or disable an eye tracker's first-person camera using a mechanical shutter [43]. However, such a method performs software-level processing on high-resolution videos acquired by traditional cameras, which may already contain privacy-sensitive data that could be exposed in an attack.

Instead of using traditional cameras to acquire the data and then using software-level processing to preserve privacy, a better idea would be to design a camera that directly obviates sensitive data while still obtaining useful information for a given task. Recently, thanks to various software and hardware advances, the entire system (camera's optical elements and image processing algorithms parameters) can be optimized in an end-to-end fashion, enabling the design

of domain-specific computational cameras [8, 24, 41, 18]. In the literature, the end-to-end optimization of domain-specific computational cameras is known as *Deep Optics*. Prior work in this line aims to improve the optical elements to acquire high-resolution/high-fidelity images and simultaneously improve the performance of computer vision algorithms. Here, we are interested in extending this philosophy to design privacy-preserving optical systems.

In this paper, we design a privacy-preserving computational camera via end-to-end optimization to capture useful information to perceive humans in the scene while hiding privacy-sensitive information. Since many computer vision applications need to analyze humans as the first step in their frameworks, we are interested in jointly optimizing a *freeform lens* (the spatially varying surface height of a lens) together with a human pose estimation (HPE) network to develop a privacy-preserving HPE system. The contributions of our work are the following: **(i)** We introduce a privacy-preserving end-to-end optimization framework to extract useful information from the scene yet preventing the imaging system from obtaining detailed and privacy-sensitive visual data. **(ii)** Using our end-to-end optimization framework, we optimize an optical encoder (Hardware-level protection) with a software decoder (convolutional neural net) to add a visual privacy protection layer to HPE. We jointly optimize the optical elements of the camera lens and fine-tune the backbone of a HPE network. We show that it is not necessary to retrain the HPE network layers to achieve privacy preservation. **(iii)** We perform extensive simulations on the COCO dataset to validate our proposed privacy-preserving deep optics approach for HPE. **(iv)** We built a proof-of-concept optical system. Our experimental results in hardware match the simulations.

In principle, our main objective is to show the benefits of a deep-optics-inspired approach to develop robust privacy-preserving vision algorithms. We design the optical system lens to degrade the image quality and obscure sensitive private information, which is opposite from the traditional approach of improving the imaging quality. We do not aim to develop a new HPE network. Instead, we add a visual privacy protection layer to an already trained HPE network using the designed optics and fine-tune the backbone layers. Our experiments show that there is a trade-off between the attained scene degradation and the HPE precision.

2. Related work

Current computer-vision algorithms for human pose estimation (HPE) do not consider privacy and rely on high-resolution images. Most existing privacy-preserving computer vision approaches tackle the action recognition task, while privacy-preserving HPE is not widely explored yet.

Human Pose Estimation. There are multiple approaches for addressing the multi-person HPE problem. Recently, convolutional networks [28, 32, 45, 46, 47, 48] have shown superior performance over prior methods such probabilistic graphical models or pictorial structures [33, 54, 55]. In general, there are two broadly used approaches for tackling the multi-person HPE problem [31]: bottom-up, where the body keypoints are predicted first and then grouped into person instances [49, 53]; and top-down, where the human bodies are detected first and then, for each detected body, joints are obtained via single-person pose estimation. Among bottom-up representative works, the OpenPose architecture [5] proposes to link the keypoints that are likely to lie in the same person using part-affinity fields. We build our privacy-preserving HPE approach on top of the OpenPose model.

Privacy-preserving Computer Vision. We divide prior work into software-level and hardware-level protection. The latter is considered more robust to attacks.

Software-level Privacy Protection. Most prior privacy-preserving computer vision methods operate after a high-fidelity image has been acquired; hence they only provide software-level privacy protection. These methods rely on domain knowledge and hand-crafted approaches, such as pixelation, blurring, and face/object replacement, to protect sensitive information [1, 9, 30]. This can be useful in practical settings when we know in advance what to protect in the scene. More recent works propose a more general approach that learns privacy-preserving encodings through adversarial training [4, 34, 52]. They actively learn to degrade or inhibit private attributes while maintaining important features to perform inference tasks. Unfortunately, there is no prior work in software-level privacy protection for HPE. The closest works study human fall detection [2] and body posture [12]. While these software-level approaches preserve privacy in the final application, the acquired images still do not protect privacy.

Hardware-level Privacy Protection. Hardware-level privacy protection approaches rely on the optical system to add a layer of security by removing sensitive data during image acquisition. Prior work uses low-resolution cameras to capture videos and avoid the unwanted leak of identity information of the human subjects [37, 38]. One can also select a defocus blur to provide a certain level of privacy over a working region within the limits of sensor size [35, 36]; however, only using optical defocus for privacy may be susceptible to reverse engineering, as we will show in Section 4. More recently, a coded aperture camera is used to directly perform human action recognition from encoded measurements without requiring image restoration as an intermediate step [51]. The only prior work on privacy-preserving HPE uses low-resolution depth images as input to an end-to-end framework that integrates a multi-scale super-resolution network with a 2D HPE network [42]. All these methods assume that the attacker has no access to the hardware. We

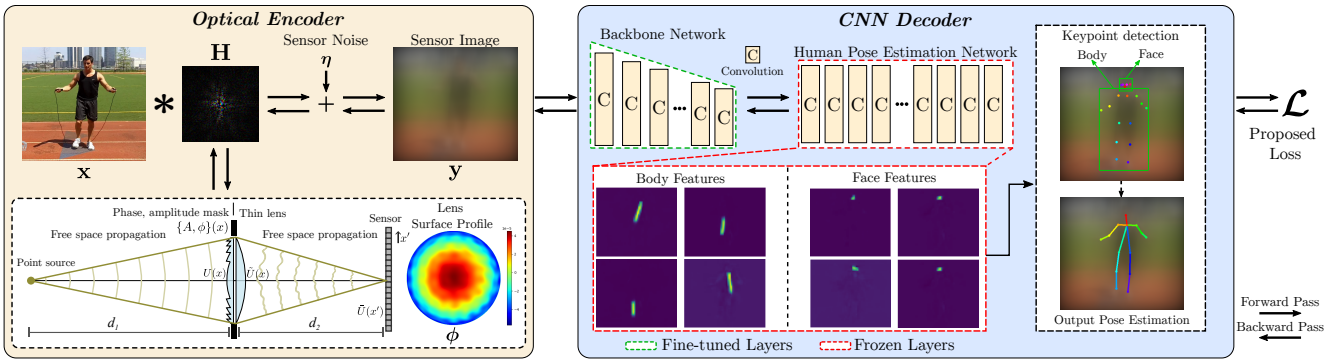

Figure 2: Our proposed end-to-end framework. The optical-encoder consists of a camera with a convex thin lens and a refractive optical element add-on. We achieve privacy-protection by jointly optimizing the optics (by adding aberrations directly on the freeform lens surface) and fine-tuning some layers of the backbone network while keeping the human pose estimation network frozen.

propose a hardware-level privacy-preserving HPE framework: we leverage *Deep optics* to design an optical lens that obscures private information while enabling HPE.

Deep Optics. Traditionally, the optics system and image processing algorithms have been developed separately. First, the optical elements are configured and fixed; and second, the parameters in the image processing algorithm are tuned to perform a specific task [26]. Recently, the idea of jointly optimizing the optical system and the image processing algorithm has drawn broad attention and is known as *Deep Optics* [41]. This idea has been successful in color imaging and demosaicing [6], extended depth of field and super-resolution imaging [41], monocular depth imaging [8, 16], image classification [7], time-off light imaging [23], high dynamic range imaging [24], and computational microscopy [17, 27]. Their philosophy is to enhance the imaging quality to improve the performance of computer vision algorithms. We introduce a radically opposite approach: we design the optical elements to degrade the image quality and obscure private information while still enabling HPE.

3. Privacy-preserving Pose Estimation

We are interested in the privacy-preserving human pose estimation task. Our general strategy is to optimize the camera optics and the human pose estimation network jointly to achieve privacy protection via image degradation. The key idea is that we can modify the camera lens to degrade the image in such a way that the identity of the subjects is obscured while preserving important features for pose estimation. To achieve this, we introduce the end-to-end framework depicted in Figure 2. Our method has two key components: an *Optical Encoder* (Section 3.1) and a *CNN Decoder* (Section 3.2). The Optical Encoder module is parametrized appropriately to allow for learning of the camera lens. The CNN Decoder performs the task of human pose estimation on our optically degraded image. During training, we optimize these two modules jointly to obtain our privacy-preserving pose estimation system. The result of the training process is two-fold: the camera lens parameters α^* and the convolutional network for pose estimation h^*. To achieve this, we can formulate a loss function for learning that combines our two goals:

$$\alpha^*, h^* = \arg\min_{\alpha,h} L_T(h) + L_P(\alpha). \quad (1)$$

where L_T is the loss function for the pose estimation task, and L_P is a loss function that encourages privacy-preservation. During inference, we can deploy our system in hardware by constructing a camera lens using the optimal parameters α^* that acquires degraded images on which our network h^* can perform pose estimation. One could also deploy a less-secure system as a software-only solution, by implementing the image degradation post-acquisition. The rest of this section describes the details of our framework.

3.1. Optical Encoder

The Optical Encoder module in Figure 2 is responsible for the image acquisition process in our privacy-preserving human pose estimation (HPE) system. As outlined earlier, our strategy for privacy-preservation is to modify the optical system of the camera during training. The goal is to produce images that visually obscure the identity of the person but still preserve important features for pose estimation. We achieve this by adopting the deep optics philosophy: we use an end-to-end training approach to jointly optimize the camera optics and the HPE network. However, our motivation diverges from prior deep optics [24, 41]: we want to optimize the camera optics by adding optical aberrations directly on the surface of the thin lens (freeform lens) instead of removing them. Furthermore, unlike prior deep optics methods, we do not perform image reconstruction and instead we work directly with the acquired low-quality image.

The key to enable such end-to-end learning is to appropriately parametrize the camera lens so that we can perform back-propagation. Note that the training signal to optimize the camera optics will back-propagate from the privacy-preserving loss $L_P(\alpha)$ (Section 3.3). There are three key parts to our parametrization: the lens surface profile ϕ, which we write in terms of Zernike coefficients α, and the

corresponding point spread function (PSF) $\mathbf{H}$ for the camera lens. First, we describe the relationship between ϕ and $\mathbf{H}$ by the image formation Model below. Then, we introduce the parametrization of the lens surface profile ϕ in terms of coefficients α for the Zernike polynomials.

Image Formation Model. We derive a wave-based image formation model for natural scenes to write the PSF $\mathbf{H}$ in terms of ϕ, assuming spatially incoherent light. Similar to recent works on end-to-end camera designs [7, 41], we model the light transport in the camera using a differentiable Fourier optics model [13].

Figure 2 depicts our optical system, which consists of a convex thin lens with a custom refractive optical element add-on with surface profile ϕ. Similar to a photographic filter, such an optical element is mounted directly in front of the lens. The response of the camera system to a point light source is described by the point spread function (PSF) created by the lens. The sensing process can be modeled as a 2D convolution operation between the scene and PSF as

$$\mathbf{y} = g(\mathbf{H} * \mathbf{x}) + \eta, \tag{2}$$

where $\mathbf{x} \in \mathbb{R}_{+}^{w \times h}$ is the scene and it is represented as a discrete color image with $w \times h$ pixels, and each pixel has value in $[0, 1]$; η represents the Gaussian noise in the sensor, and $g(\cdot)$ is the camera response function, which we assume linear. This model also assumes that the PSF is shift-invariant, but the model could be generalized.

Assuming that the thin lens has a focal length f at a distance d_2 from the sensor, the relationship between the in-focus distance and the sensor distance in the paraxial ray approximation is given by the thin-lens equation: $1/f = 1/d_1 + 1/d_2$. Therefore, an object at a distance d_1 in front of the lens appears in focus at a distance d_2 behind the lens. Assuming that the scene is at optical infinity, we first propagate the light emitted by the point, represented as a spherical wave, to the lens. The complex-valued wave field immediately before the lens is given by:

$$U(x, y) = \exp\left(ik\sqrt{x^2 + y^2 + z^2}\right), \tag{3}$$

where $k = 2\pi/\lambda$ is the wavenumber. The refractive optical element first delays the phase of this incident wavefront by an amount proportional to the surface profile $\boldsymbol{\phi}$ of the optical element at each point (x, y). Equivalently, the optical element may be represented by a multiplicative phase transformation of the form

$$t_{\boldsymbol{\phi}}(x, y) = \exp(ik(n(\lambda) - 1)\phi(x, y)), \tag{4}$$

where $n(\lambda)$ is the wavelength-dependent refractive index of the optical element material.

The light wave continues to propagate to the camera lens, which induces the following phase transformation [13]

$$t_l(x, y) = \exp\left(-i\frac{k}{2f}(x^2 + y^2)\right). \tag{5}$$

Considering that a lens has a finite aperture size, we use a binary circular mask $A(x, y)$ with diameter D to model the aperture and block light in regions outside the open aperture. To find the electric field immediately after the lens, we multiply the amplitude and phase modulations of the refractive optical element and lens with the input electric field:

$$\tilde{U}(x, y) = A(x, y)t_{\phi}t_l(x, y)U(x, y). \tag{6}$$

Finally, the field propagates a distance d_2 to the sensor with the exact transfer function [13]:

$$T_{d_2}(f_x, f_y) = \exp\left[ikd_2\sqrt{1 - (\lambda f_x)^2 - (\lambda f_y)^2}\right], \tag{7}$$

where (f_x, f_y) are spatial frequencies. This transfer function is applied in the Fourier domain as:

$$\bar{U}(x', y') = \mathcal{F}^{-1}\left\{\mathcal{F}\left\{\tilde{U}(x, y)\right\} \cdot T_{d_2}(f_x, f_y)\right\}, \tag{8}$$

where $\mathcal{F}$ denotes the 2D Fourier transform. Since the sensor measures light intensity, we take the magnitude-squared to find the values of the PSF $\mathbf{H}$ at each position (x, y) as:

$$H(x', y') = |\bar{U}(x', y')|^2. \tag{9}$$

Lens Parametrization. We parametrize the lens surface profile $\boldsymbol{\phi}$ with the Zernike basis, which leads to smoother surfaces, as

$$\boldsymbol{\phi} = \sum_{j=1}^{q} \alpha_j \mathbf{Z}_j, \tag{10}$$

where $\mathbf{Z}_j$ is the j-th Zernike polynomial in Noll notation, and α_j is the corresponding coefficient [3]. Each Zernike polynomial describes a wavefront aberration; hence the surface profile $\boldsymbol{\phi}$ is formed by the linear combination of all aberrations. In this regard, the optical element parameterized by $\boldsymbol{\phi}$ can be seen as an optical encoder, where the coefficients α_j determine the data transformation. Therefore, different from common deep optics approaches, our end-to-end training finds a set of coefficients $\alpha^* = \{\alpha_j\}_{j=1}^{q}$ that provides the maximum visual distortion of the scene but allows to extract relevant features to perform HPE.

3.2. CNN Decoder

To perform HPE, we use the OpenPose network architecture [5]. The OpenPose network is composed of a VGG-19 [40] backbone, and two branches of convolutional layers. The backbone network extracts features from an image of size $w \times h$, which are then fed into the two branches. One branch predicts a set of confidence maps, where each map represents a specific body part location; the second branch predicts a set of Part Affinity Fields (PAFs), where each field represents the degree of association between parts. Successive stages are performed to refine the predictions made by each branch. Finally, the confidence maps and the PAFs are parsed by greedy inference to produce the 2D locations of body keypoints for each person in the image [5].

HPE Loss Function L_T. The OpenPose loss accounts for both the body and face keypoints to improve human pose estimation in an image. Let $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_E\}$ be the set of confidence maps, where each map $\mathbf{S}_e \in \mathbb{R}^{w \times h}$ represents a specific keypoint location, $e \in \{1, \cdots, E\}$. Similarly, let $\mathcal{V} = \{\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_C\}$ be the set of PAFs, where each affinity field $\mathbf{V}_c \in \mathbb{R}^{w \times h \times 2}$ represents the degree of association between the keypoints, and $c \in \{1, \cdots, C\}$. We split the confidence maps as $\mathcal{S} = \{\mathcal{S}_B, \mathcal{S}_F\}$, where $\mathcal{S}_B \subseteq \mathcal{S}$ contains the maps $\mathbf{S}_e$ for the body keypoints, and

$\mathcal{S}_F \subseteq \mathbf{S}$ contains the maps $\mathbf{S}_e$ for the face keypoint locations. Similarly, we split the PAFs as $\mathcal{V} = \{\mathcal{V}_B, \mathcal{V}_F\}$, where $\mathcal{V}_B \subseteq \mathcal{V}$ contains the affinity fields $\mathbf{V}_c$ of the body limbs, and $\mathcal{V}_F \subseteq \mathcal{V}$ contains the affinity fields $\mathbf{V}_c$ that represent the degree of association between two face parts. We define the loss function for a subset of keypoints χ at stage τ is

$$F_\tau(\chi) = \sum_{\delta=1}^{|\chi|} \sum_{\mathbf{p}} \mathbf{B}(\mathbf{p}) \cdot \|\chi_\delta^\tau(\mathbf{p}) - \chi_\delta^*(\mathbf{p})\|_2^2 \quad (11)$$

where $|\chi|$ is the number of keypoints in the subset. For instance, if $\chi = \mathcal{S}_B$ then $|\chi|$ will be the total number of body-related confidence maps. $\mathbf{B}$ is a binary mask with $\mathbf{B}(p) = 0$ when the annotation is missing at the pixel p, and χ_δ^* denotes the groundtruth. Then, the overall OpenPose L_T is

$$L_T = \sum_{\tau=1}^{\Gamma_1} F_\tau(\mathcal{V}_B) + F_\tau(\mathcal{V}_F) + \sum_{\tau=\Gamma_1+1}^{\Gamma_1+\Gamma_2} F_\tau(\mathcal{S}_B) + F_\tau(\mathcal{S}_F), \quad (12)$$

where Γ_1 and Γ_2 denote the total of PAF and confidence map stages, respectively.

3.3. Privacy-preserving Loss Function L_P

Defining a privacy-preserving loss function is not a straightforward task, and the definition will depend on concrete application contexts. There are various privacy-related attributes, such as the face, race, gender, or age [29]. However, the face is the main attribute we would like to obscure in our privacy-preserving vision task. Therefore, we define the privacy-preserving loss taking into account the face keypoints detection in the images. In principle, we are not interested in obtaining an accurate localization of face keypoints, and we would like to obscure such face regions from the image. Then, we only want to preserve the body keypoints and let the end-to-end training degrade all the image's spatial details (including the faces). To further enforce image degradation, we maximize the ℓ_2 norm error between the original image $\mathbf{x}$ and the acquired image $\mathbf{y}$, defined as

$$L_E = \sum_b \frac{1}{2} \|\mathbf{y}_b - \mathbf{x}_b\|_2^2, \quad (13)$$

where the subscript b denotes the color bands of the RGB images. We define the privacy-preserving loss function as

$$L_P = -\sum_{\tau=1}^{\Gamma_1} F_\tau(\mathcal{V}_F) - \sum_{\tau=\Gamma_1+1}^{\Gamma_1+\Gamma_2} F_\tau(\mathcal{S}_F) - \beta_2 L_E. \quad (14)$$

Finally, considering Eq. 1, we compute the total loss at the end of our proposed framework, as follows

$$L = L_T + L_P = \beta_1 \Big(\sum_{\tau=1}^{\Gamma_1} F_\tau(\mathcal{V}_B) + \sum_{\tau=\Gamma_1+1}^{\Gamma_1+\Gamma_2} F_\tau(\mathcal{S}_B) \Big) - \beta_2 L_E. \quad (15)$$

3.4. Training Details

Optics Layer Simulation. We simulate a sensor with pixel size of 3.40μm and resolution of 864×864 pixels. We use the first $q = 350$ Zernike coefficients in Noll notation to shape the surface profile ϕ. The fourth Zernike coefficient (the defocus term) is initialized, such that the lens has a focal length of $f = 25$mm. The optical element is discretized with a 3.40μm feature size on an 864×864 grid.

Fine-tuning. We are interested in adding a privacy protection layer to a pre-trained OpenPose network. Hence, to perform training we assume an aberration-free freeform lens and use the pretrained weights of a Tensorflow implementation of OpenPose [22] as a starting point. After initialization with the pre-trained weights, we freeze the two branches of OpenPose and only fine-tune some layers of the VGG-19 backbone with a lower learning rate to learn extracting human body features from the private image $\mathbf{y}$. Figure 2 illustrates the frozen and fine-tuned layers.

Training. During training, we first perform one forward pass through the network by convolving the images from the training set with the PSF $\mathbf{H}$ to obtain the optically-encoded sensor image $\mathbf{y}$, as described by Eq. 2. Next, the VGG-19 backbone extracts features from $\mathbf{y}$, and then the features are fed into the two branches of the OpenPose architecture. Now, we split the confidence maps $\mathcal{S}$ and PAFs $\mathcal{V}$ into body-related and face-related features as described in the Section 3.2, and compute the loss described in Eq. 15. After computing L, we use the automatic differentiation capabilities of Tensorflow to back-propagate the error and update the parameters of the VGG-19 backbone and the coefficients α_j that model surface profile ϕ of the lens using Eq. 10. We trained the end-to-end model using Adam optimizer with a batch size of 22 and an initial learning rate of 2×10^{-5}. We applied an exponential learning rate decay with a decay factor of 0.666 that is triggered after 15K, 20K, 25K, 28K, and 35K training steps. We trained the network for 50K steps (gradient updates), which took about 24 hours on a Tesla V100-SXM2 GPU with 32 GB of memory.

4. Experimental Results

The goal of our work is privacy-preserving pose estimation, so evaluate performance in the task of human pose estimation (HPE), as well as the level of privacy protection. We evaluate HPE following standard practice. To evaluate privacy protection, we use two indirect proxies: image degradation and face recognition. Our experiments are performed on two implementations of our framework: a software-only simulation and a hardware prototype built in the lab.

Dataset, Metrics and Evaluation Method. We train our proposed end-to-end approach on the COCO [21] 2017 keypoints dataset and evaluate our approach on the val2017 set. To quantitatively evaluate HPE, we use the standard COCO evaluation metric: Object Keypoint Similarity (OKS) [21]. Since we aim at preserving privacy, we expect the estimation of face keypoints to degrade, while we want to maintain good performance on the estimation of body keypoints. To make a fair comparison, we sightly modify the COCO evaluation script to not consider the face keypoints. We report the standard average precision (AP) and recall (AR) scores: AP, AP^{50}, AP^{75}, AP^M (medium objects), AP^L (large objects), and AR. To measure image

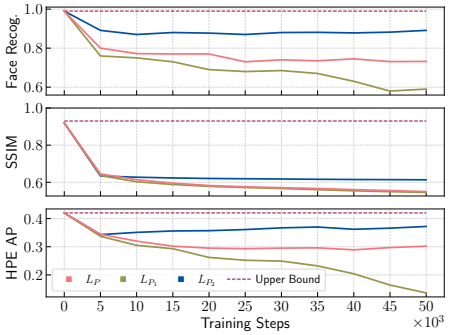

Figure 3: Experiment I. Comparison of different privacy-preserving losses. The performance of the proposed end-to-end framework using different losses is depicted with different colors.

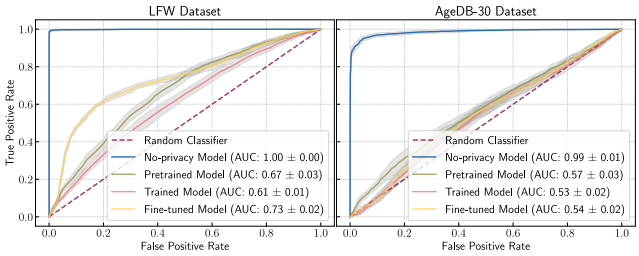

Figure 4: Experiment II. Face recognition performance on images acquired with our optimized lens.

degradation, we use the peak-signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) [19]. Large values of PSNR or SSIM indicate high quality. Thus, we expect to achieve the minimum PSNR or SSIM values while achieving high AP on the human pose keypoints. We report the average PSNR and SSIM over all images from the validation set. Finally, we use an implementation of the face recognition network ArcFace [10] to measure privacy. We train ArcFace on Microsoft Celeb (MS-Celeb-1M) [14] and test on LFW [20], AgeDB-30[25] and CFP-FP [39]. We measure face recognition performance in terms of the area under curve (AUC) of the ROC curve.

4.1. Simulation Experiments

Ablation study. We conduct five ablation experiments and investigate different configurations for our architecture.

Experiment I explores two alternative formulations for the privacy-preserving loss L_P. We define such losses as:

$$L_{P_1} = -\beta_3 \Big(\sum_{\tau=1}^{\Gamma_1} F_\tau(\mathcal{V}_F) - \sum_{\tau=\Gamma_1+1}^{\Gamma_1+\Gamma_2} F_\tau(\mathcal{S}_F) \Big) - \beta_2 L_E \quad (16)$$

$$L_{P_2} = - \sum_{\tau=1}^{\Gamma_1} F_\tau(\mathcal{V}_F) - \sum_{\tau=\Gamma_1+1}^{\Gamma_1+\Gamma_2} F_\tau(\mathcal{S}_F) - \beta_4 L_F, \quad (17)$$

where $\beta_3 > 1$,

$$L_F = Sim_{cos}(a_f(\mathbf{x}), a_f(\mathbf{y})), \quad (18)$$

Sim_{cos} denotes the cosine similarity, and $a_f(\cdot)$ stand for the ArcFace model [10]. To compute L_F, we use the pretrained ArcFace model on faces extracted from the input image $\mathbf{x}$ and distorted image $\mathbf{y}$. See Section 2 of supplementary material for more details. Figure 3 shows performance versus training step obtained by training with each of the three privacy-preserving losses (each in a different color). Performance is measured under three viewpoints: Face recognition AUC, HPE AP, and image degradation SSIM. We show with dashed lines the upper bound for each metric. Face recognition AUC is calculated on the LFW dataset. To do this, for each privacy-preserving HPE model trained with a specific privacy loss, we first generate a "private" MS-Celeb-1M dataset and train the ArcFace model – this equivalent to an attack that can obtain an annotated set of face images acquired with our camera. We do not train ArcFace from scratch; instead, we load the pretrained weights and fine-tune the model using the "private" set. We observe that L_{P_1} produces highly distorted images, and the face recognition performance is poor; however, the HPE AP is the lowest in comparison with the other losses. L_{P_2} obtains the best HPE performance, but image distortion decreases slowly and seems to stabilize after 25K training steps; hence the face recognition achieves good performance, which is not desired. Our proposed privacy-preserving loss L_P achieves good HPE results and low face recognition performance.

In *Experiment II*, we test the face recognition performance on images acquired using our proposed privacy-preserving lens on the LFW and AgeDB-30 datasets. Figure 4 show the ROC curves for each testing approach: "No-privacy Model" uses the pretrained ArcFace model on the original images; "Pretrained model" uses the pretrained ArcFace model on the private version of each dataset; "Trained model" uses an ArcFace model trained from scratch using the private version of the MS-Celeb-1M dataset; "Fine-tuned Model" uses a pretrained ArcFace model fine-tuned with the private version of the MS-Celeb-1M dataset. As observed, the fine-tuned model performs best on the LFW dataset compared to the other testing approaches. However, the ArcFace model performance is similar to a random classifier on the AgeDB-30 dataset for all the testing approaches. The ArcFace model does not perform well on the images generated by our proposed lens design. See results with the CFP-FP dataset in our supplementary.

In *Experiment III*, we optimize for $q = 350$ Zernike coefficients and do not fine-tune the HPE backbone layers. Table 1 shows that the optimization did not converge to an optimal point; hence the network is unable to estimate the pose in the degraded image even after few training steps. After training, we obtain a mean PSNR of 11.452 and SSIM of 0.496. In *experiment IV*, we fix the number of Zernike coefficients to $q = 350$ and fine-tune the first 10,20, and 40 layers of the VGG-19 backbone. Fine-tuning the first 20 layers leads to the best results in AP while achieving low PSNR and SSIM values. In *experiment V*, we fix the layers to be fine-tuned in the network and train using a different

Experiment	Fine-tuned Layers	Zernike Coefficients	PSNR	SSIM	AP
III	No Fine-tune	350	11.452	0.496	-
IV	10	350	14.598	0.565	0.263
	20	**350**	**14.851**	**0.567**	**0.302**
	40	350	14.577	0.562	0.251
V	20	15	16.692	0.582	0.168
	20	50	16.328	0.579	0.231
	20	150	16.142	0.571	0.258

Table 1: Ablation study of our method on COCO val2017 dataset using the OpenPose network. The configuration shown in bold leads to the best results in terms of image degradation and AP.

Method	PSNR	SSIM	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
OPPS [5]	-	-	0.421	0.655	0.439	0.444	0.428	0.506
Defocus Lens[35]	16.614	0.598	0.197	0.432	0.155	0.126	0.299	0.256
Low-Resolution[38]	18.54	0.476	0.067	0.197	0.032	0.031	0.123	0.106
PP-OPPS (Ours)	**14.851**	**0.567**	**0.302**	**0.555**	**0.266**	**0.276**	**0.359**	**0.363**

Table 2: Comparisons on the COCO validation set. We compare our method against two traditional privacy-preserving approaches: Defocus and Low-resolution cameras. The PP prefix stands for our proposed privacy-preserving approach.

number of Zernike coefficients. Increasing the number of Zernike coefficients leads to better encoding; hence AP increases while PSNR and SSIM decrease. However, memory consumption also increases linearly since we need to store all the Zernike basis. In the following experiments, we use the best configuration from Table 1 (shown in bold font).

Comparison with other methods. Privacy-preserving HPE methods are not well explored in the literature. Therefore, to compare our method, we adapt the ideas of using low-resolution cameras [38] and cameras with a defocus lens [35] to provide visual privacy protection. We simulate both types of cameras, fix the optics so that the lens is not optimized during learning, and fine-tune the first 20 trainable layers of the HPE backbone network similarly to our proposed approach. To implement the low-resolution approach, we use images with a resolution of 32×32. We compare our method with a Tensorflow implementation of OpenPose (OPPS) [5] architecture [22]. Table 2 reports the COCO keypoints evaluation results and the average of the PSNR and SSIM image quality metrics among all images from the COCO validation set. In the table, PP-OPPS stands for our proposed privacy-preserving approach for OPPS. The simple defocus lens achieves an AP of 0.197 on low-quality images with an average PSNR of 16.614 and SSIM of 0.598. Our proposed optimized lens leads to better results since it adds more aberrations to the optical system than the defocus lens approach, which only incorporates one aberration (defocus). The low-resolution approach does not work well with our proposed end-to-end training approach since it leads to the lower average SSIM value and the lowest performance in terms of keypoints AP. See our supplementary for results when using other HPE network.

Robustness to Deconvolution. We investigate the robustness of our proposed lens design to deconvolution attacks. In the worst scenario, we assume that an attacker knows the set of Zernike coefficients that form the surface profile ϕ, i.e., the PSF is known. Then, the attacker could perform a non-blind deconvolution to reveal the identity of a person within the scene. Figure 6 illustrates our results. Although the defocus lens seems to obscure visual details, it is susceptible to reverse engineering, and the identity of people can be revealed using Wiener deconvolution [11]. The deconvolution approach does not work well for our proposed lens design as it has significantly more aberrations, making it more robust. In a more realistic scenario, an attacker can access a large collection of blur images acquired with our proposed camera but does not know the PSF. We already explore this scenario (blind-deconvolution) and present some results in Section 5 of our supplementary.

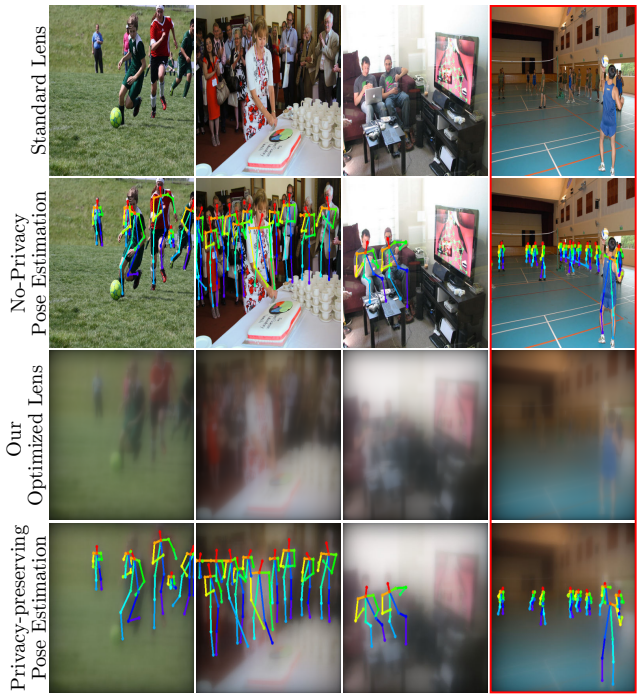

Figure 5: Qualitative results on example COCO images. We compare our proposed privacy-preserving pose estimation results using the optimized lens with the Non-privacy approach using a standard lens. The last column depicts a failure case where we fail to estimate the pose of far distant people.

Qualitative results. Figure 5 shows a visual comparison of our proposed method using the optimized lens against the results from the original OpenPose (No-privacy pose estimation), which works on images acquired with a standard lens. Our proposed privacy-preserving approach achieves good human pose estimation on degraded images. The last column shows an example failure case of our method; as observed, the method fails to estimate distant people's pose. However, when a person is far from the camera, less is the privacy concern; hence the privacy protection given by our method is still useful in most cases.

Figure 6: Non-blind deconvolution of private images acquired with a defocusing lens compared to our lens. Our image is more robust to deconvolution even when the PSF is know.

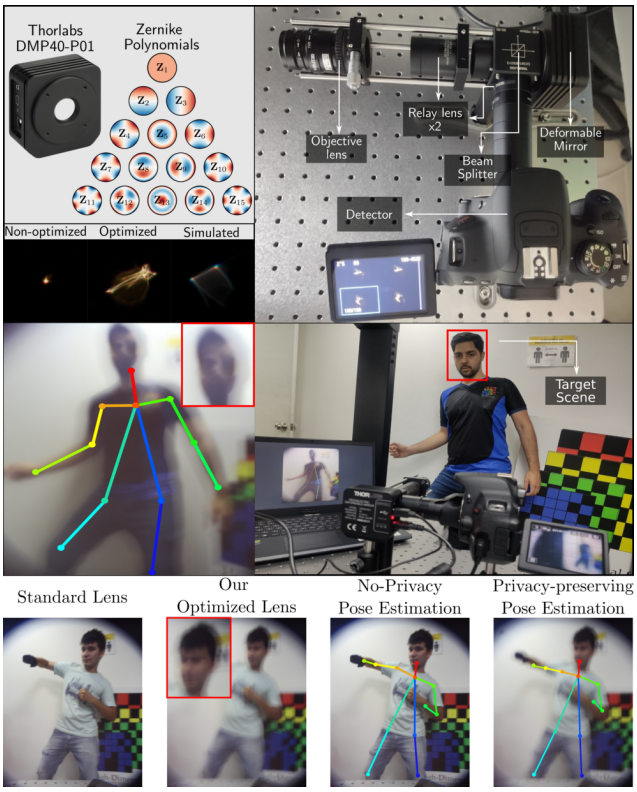

Figure 7: (Top) Experimental Hardware Setup for our privacy-preserving approach. (Bottom) Qualitative results on some example images acquired by the prototype camera.

4.2. Hardware Experiments

To experimentally evaluate the effectiveness of our proposed privacy-preserving approach, we built the proof-of-concept optical system in Figure 7. The prototype comprises a main objective lens coupled with a $4f$ system, which has a phase modulating element at $2f$. Our Camera is a CANON EOS REBEL T5i placed at the optical setup's image plane. The intermediate image plane is formed by an 8mm objective lens (NAVITAR MVL8M23), which is relayed by a pair of 75mm Fourier transforming lenses (Thorlabs AC254-075-A-ML). Using a beamsplitter (BS, Thorlabs CCM1-BS013), we placed a deformable mirror (DM, Thorlabs DMP40-P01) at the pupil plane at a distance of $2f = 150$mm from the intermediate image plane. Finally, the camera detector is placed at a distance of $2f = 150$mm from the deformable mirror. We captured a series of images of a point white light source using a pinhole of 20 μm to calibrate the acquired PSFs.

Total Images	Fine-tuning set	Testing set	AP	AP^{50}	AP^{75}	AR
300	150	150	0,562	0,731	0.532	0.584

Table 3: Quantitative evaluation of acquired images in our Lab.

After calibrating the system, we obtain a Non-optimized PSF, i.e., we have an optical system that captures non-privacy RGB images. Then, we simulate the system using our proposed approach to obtain the optimized Zernike polynomials and setup the deformable mirror. The deformable mirror uses the Zernike polynomials to deform its surface, thus modifying the incident light wavefront. The optical system with the deformable mirror creates an optimized PSF that closely resembles the simulated PSF and captures private images, see Fig. 7. We use a small set of captured measurements to fine-tune the HPE network for a few epochs.Finally, we run human pose estimation on our images using the trained privacy-preserving HPE network. Figure 7 shows predicted poses on the acquired private and non-private images. Table 3 presents a quantitative evaluation on a small testing set captured in the laboratory.

Limitations. The deformable mirror is the main limitation of the proof-of-concept optical system. This device can only use $q = 15$ Zernike Polynomials, which limits the level of distortion of the scene. However, results show that our acquired images successfully protect personal identity by distorting the face. We also performed a quantitative evaluation on a small set of images acquired in the lab. Due to pandemic restrictions, we cannot acquire a larger scale image dataset in the lab. For now, our small scale tests show results consistent with our extensive experiments.

5. Conclusion

We presented a privacy-preserving approach for pose estimation that consists of an optical encoder that obscures sensitive private information and a decoder that performs HPE on degraded images. We extensively evaluated and experimentally validated our approach on simulations and a hardware prototype. Our qualitative and quantitative results indicate a trade-off between image degradation and HPE accuracy. We plan to extend our method to other computer vision tasks. We will also consider more complex PSFs, such as depth-invariant PSFs, which may improve the HPE of far distant people.

References

[1] Prachi Agrawal and PJ Narayanan. Person de-identification in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(3):299–310, 2011. 2

[2] Umar Asif, Benjamin Mashford, Stefan Von Cavallar, Shivanthan Yohanandan, Subhrajit Roy, Jianbin Tang, and Stefan Harrer. Privacy preserving human fall detection using video data. In *Machine Learning for Health Workshop*, 2020. 2

[3] Max Born and Emil Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013. 4

[4] Karla Brkic, Ivan Sikiric, Tomislav Hrkac, and Zoran Kalafatic. I know that person: Generative full body and face de-identification of people in images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. 2

[5] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser A Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 4, 7

[6] Ayan Chakrabarti. Learning sensor multiplexing design through back-propagation. In *Advances in Neural Information Processing Systems*, 2016. 3

[7] Julie Chang, Vincent Sitzmann, Xiong Dun, Wolfgang Heidrich, and Gordon Wetzstein. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Scientific reports*, 8(1):1–10, 2018. 3, 4

[8] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2, 3

[9] Datong Chen, Yi Chang, Rong Yan, and Jie Yang. Tools for protecting the privacy of specific individuals in video. *EURASIP Journal on Advances in Signal Processing*, 2007:1–9, 2007. 2

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 6

[11] Jiangxin Dong, Stefan Roth, and Bernt Schiele. Deep wiener deconvolution: Wiener meets deep learning for image deblurring. In *Advances in Neural Information Processing Systems*, 2020. 7

[12] Munkhjargal Gochoo, Tan-Hsu Tan, Fady Alnajjar, Jun-Wei Hsieh, and Ping-Yang Chen. Lownet: Privacy preserved ultra-low resolution posture image classification. In *IEEE International Conference on Image Processing (ICIP)*, 2020. 2

[13] Joseph W Goodman. *Introduction to Fourier optics*. Macmillan Learning, 4 edition, 2017. 4

[14] Yandong Guo, Lei Zhang, Yuxiao Hu, X. He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 6

[15] Albert Haque, Arnold Milstein, and Li Fei-Fei. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature*, 585(7824):193–202, 2020. 1

[16] Lei He, Guanghui Wang, and Zhanyi Hu. Learning depth from single images with deep neural network embedding focal length. *IEEE Transactions on Image Processing*, 27(9):4676–4689, 2018. 3

[17] Eran Hershko, Lucien E Weiss, Tomer Michaeli, and Yoav Shechtman. Multicolor localization microscopy and point-spread-function engineering by deep learning. *Optics express*, 27(5):6158–6183, 2019. 3

[18] Carlos Hinojosa, Jorge Bacca, and Henry Arguello. Coded aperture design for compressive spectral subspace clustering. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1589–1600, 2018. 2

[19] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*. IEEE, 2010. 6

[20] Gary B. Huang, Marwan Mattar, Honglak Lee, and Erik Learned-Miller. Learning to align from scratch. In *NIPS*, 2012. 6

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 2014. 5

[22] Jiawei Liu, Yixiao Guo, Guo Li, Luo Mai, and Dong Hao. Hyperpose: Real-time human pose estimation. https://github.com/tensorlayer/hyperpose, 2020. 5, 7

[23] Morgan McGuire, Wojciech Matusik, Hanspeter Pfister, Billy Chen, John F Hughes, and Shree K Nayar. Optical splitting trees for high-precision monocular imaging. *IEEE Computer Graphics and Applications*, 27(2):32–42, 2007. 3

[24] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3

[25] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017. 6

[26] Shree K Nayar. Computational cameras: Redefining the image. *Computer*, 39(8):30–38, 2006. 3

[27] Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E Weiss, Onit Alalouf, Tal Naor, Reut Orange, Tomer Michaeli, and Yoav Shechtman. Deepstorm3d: dense 3d localization microscopy and psf design by deep learning. *Nature Methods*, 17(7):734–740, 2020. 3

[28] Xuecheng Nie, Jiashi Feng, Yiming Zuo, and Shuicheng Yan. Human pose estimation with parsing induced learner. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[29] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 5

[30] José Ramón Padilla-López, Alexandros Andre Chaaraoui, and Francisco Flórez-Revuelta. Visual privacy protection methods: A survey. *Expert Systems with Applications*, 42(9):4177–4195, 2015. 2

[31] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[32] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[33] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2

[34] Francesco Pittaluga, Sanjeev Koppal, and Ayan Chakrabarti. Learning privacy preserving encodings through adversarial training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019. 2

[35] Francesco Pittaluga and Sanjeev J Koppal. Privacy preserving optics for miniature vision sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2, 7

[36] Francesco Pittaluga and Sanjeev Jagannatha Koppal. Pre-capture privacy for small vision sensors. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2215–2226, 2016. 2

[37] Michael Ryoo, Kiyoon Kim, and Hyun Yang. Extreme low resolution activity recognition with multi-siamese embedding learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[38] Michael S. Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2, 7

[39] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016. 6

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[41] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM TOG*, 2018. 2, 3, 4

[42] Vinkle Srivastav, Afshin Gangi, and Nicolas Padoy. Human pose estimation on privacy-preserving low-resolution depth images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019. 2

[43] Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling. Privaceye: privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019. 1

[44] Benyue Su, Huang Wu, Min Sheng, and Chuansheng Shen. Accurate hierarchical human actions recognition from kinect skeleton data. *IEEE Access*, 7:52532–52541, 2019. 1

[45] Ke Sun, Cuiling Lan, Junliang Xing, Wenjun Zeng, Dong Liu, and Jingdong Wang. Human pose estimation using global and local normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[46] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019. 2

[47] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014. 2

[48] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[49] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[50] Lei Wang, Du Q Huynh, and Piotr Koniusz. A comparative review of recent kinect-based action recognition algorithms. *IEEE Transactions on Image Processing*, 29:15–28, 2019. 1

[51] Zihao W Wang, Vibhav Vineet, Francesco Pittaluga, Sudipta N Sinha, Oliver Cossairt, and Sing Bing Kang. Privacy-preserving action recognition using coded aperture videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 2

[52] Zhenyu Wu, Haotao Wang, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[53] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 2

[54] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR 2011*, 2011. 2

[55] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010. 2