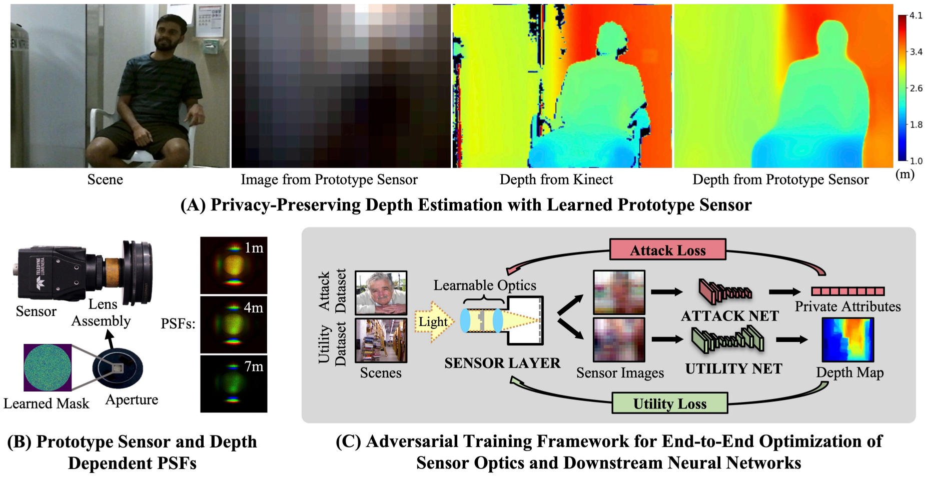
Learning Phase Mask for Privacy-Preserving Passive Depth Estimation

Zaid Tasneem[2], Giovanni Milione[1], Yi-Hsuan Tsai[1], Xiang Yu[1], Ashok Veeraraghavan[2], Manmohan Chandraker[1,3], and Francesco Pittaluga[1]

[1] NEC Laboratories America
[2] Rice University
[3] University of California San Diego



(A) Privacy-Preserving Depth Estimation with Learned Prototype Sensor

(B) Prototype Sensor and Depth Dependent PSFs

(C) Adversarial Training Framework for End-to-End Optimization of Sensor Optics and Downstream Neural Networks

Fig. 1. Learning Phase Mask for Privacy-Preserving Passive Depth Estimation

Abstract. With over a billion sold each year, cameras are not only becoming ubiquitous, but are driving progress in a wide range of domains such as mixed reality, robotics, and more. However, severe concerns regarding the privacy implications of camera-based solutions currently limit the range of environments where cameras can be deployed. The key question we address is: Can cameras be enhanced with a scalable solution to preserve users' privacy without degrading their machine intelligence capabilities? Our solution is a novel end-to-end adversarial learning pipeline in which a phase mask placed at the aperture plane of a camera is jointly optimized with respect to privacy and utility objectives. We conduct an extensive design space analysis to determine operating points with desirable privacy-utility tradeoffs that are also amenable to sensor fabrication and real-world constraints. We demonstrate the first working prototype that enables passive depth estimation while inhibiting face identification.

Keywords: Privacy, Optics, Deep Learning, Adversarial Training

1 Introduction

Computer vision is increasingly enabling automatic extraction of task-specific insights from images, but its use in ubiquitously deployed cameras poses significant privacy concerns [3, 35]. These concerns are further heightened by the fact that most cameras today are connected to the internet, leaving them vulnerable to data sniffing attacks. Existing solutions for improving visual privacy include post-capture image sanitization (blurring, resolution loss, etc.), or post-capture image encryption. Unfortunately, these solutions are vulnerable to typical sniffing attacks that can get direct access to the original captures rich in sensitive information.

This leads to two fundamental questions: Can computational cameras for machine intelligence be designed to excel at particular tasks while ensuring *pre-capture privacy* with respect to specific sensitive information? And, can such cameras be realized in practice, to achieve advantageous privacy-utility tradeoffs despite non-idealities in the modeling and fabrication process? This paper answers both the questions in the affirmative.

Unlike conventional post-capture visual privacy methods, our learned camera optics filter out privacy sensitive information prior to image capture, directly from the incident light-field – ensuring that private data never reaches the digital domain where it's susceptible to sniffing and other attacks. The learned filter can be optically implemented using a single phase mask inserted in the aperture plane of a normal camera – making the design practical and scalable. Development of our pre-capture privacy aware computational camera is driven by a novel adversarial-learning-based design principle for jointly optimizing the phase mask and downstream neural networks that enables us to achieve flexible tradeoffs for the utilitiy task of depth estimation and privacy task of avoiding face recognition.

Since, many downstream computer vision applications such as scene understanding, action recognition, planning and navigation have depth as a prerequisite, it serves as an information rich utility objective. We validate this notion by showing depth-based action recognition using private depth estimates from our learned sensor. Our optimized phase masks filter out high frequency information to obfuscate face identity while the resulting depth dependent encodings enable depth estimation.

We make the following contributions:

1. We present an end-to-end adversarial learning framework for optimizing the sensor optics with respect to utility and privacy objectives. We demonstrate the application of this design principle by optimizing the phase mask of a sensor to enable depth estimation while inhibiting face recognition.
2. We conduct an extensive design space analysis of sensor configurations with respect to phase mask design, focus settings and resolution to determine operating points with desirable privacy-utility tradeoffs that are also amenable to sensor fabrication and real-world constraints.

3. We study the impact of undiffracted light and demonstrate that it plays an important role when designing diffractive optics for privacy filtering.
4. We present the first physically realized learned computational camera that has been shown, via quantitative evaluations on real data, to provide pre-capture privacy and high utility. Through both simulated and real world experiments, we demonstrate that our prototype successfully renders human faces unidentifiable while enabling estimation of depth maps.

2 Related Work

The intersection of privacy, computer vision and computational imaging is related to many research areas. Here we summarize related works and explain how our framework is distinct. With respect to privacy, the focus of this paper is on visual privacy [37], i.e., on inhibiting estimation of sensitive attributes from imagery data. Other forms of privacy, such differential privacy [14] and federated learning [63] that aim to publish aggregate information about a database while limiting disclosure of database records are outside the scope of this paper.

In recent years, concerns regarding data sniffing attacks have led to development of pre-capture privacy cameras that apply privacy encodings at the sensor level via a trusted-hardware layer [56, 15, 42, 34] and/or filtering optics [41, 40, 36, 55]. Our approach, like [49, 20], is a generalization of these methods in that de-identification is driven through automatic inference rather than manually designed strategies. However, we employ a more realistic physics model, which enables us to reproduce our simulated results in a fully working hardware prototype. To our knowledge, we are the first to successfully port a learned pre-capture privacy sensor to a real prototype device and demonstrate, via quantitative evaluations on real data, that it provides both privacy and utility. Finally, we also show an in-depth analysis of the sensor design space that provides critical insight into how privacy is being achieved.

Many prior visual privacy methods have relied on domain knowledge and hand-crafted heuristics—such as pixelation, blurring, face replacement, etc.—to degrade sensitive information [37]. Such approaches usually fail to achieve privacy-utility trade-offs comparable to that of more recent learning-based visual privacy methods. The most successful learning-based visual privacy methods [58, 9, 39, 61], leverage adversarial training to learn encoding functions that inhibit estimation of private attributes by downstream discriminator models, yet still enable estimation of utility attributes by downstream utility models. This is a natural formulation, as an effective attack method to estimate the value of a private attribute is to train a neural network on a large set of encoded images. A similar approach has also been used to learn encoding functions that produce fair or unbiased encodings [24, 2, 25]. Such encodings can be thought of as private representations invariant to sensitive attributes such as ethnicity or gender. Adversarial training has also been used to learn adversarial perturbations that fool classifiers that expect natural images, but such classifiers recover when retrained on examples with the perturbations [31, 32, 10, 45]. Finally, [38] learns adversarial perturbations for specific camera camera optics and image processing pipelines. We seek image transformations that inhibit estimation of the private attributes even after a classifier is retrained on encoded images.

Recent works have shown that modern deep learning tools can be used to efficiently model and optimize the end-to-end computational imaging process. This approach has been successfully leveraged to design computational sensors with improved performance across a range of tasks: demoisaicing [7], monocular depth estimation [19, 18, 57, 8, 4], extended depth of field and super-resolution [47], non-paraxial imaging [23], object detection [51] and high dynamic range imaging [48, 30]. We present a computational imaging design principle that not only enables improved performance on a target utility task, but also inhibits estimation of private attributes.

3 Method

Our goal is end-to-end optimization of a sensor's optical elements with respect to privacy and utility objectives. To achieve this, we employ an adversarial learning formulation in which a sensor layer with learnable parameters is trained to simultaneously (a) promote the success of UtilityNet, a downstream neural network aims to solve a target vision task, e.g., depth estimation, and (b) inhibit the success of AttackNet, a downstream neural network that seeks to infer private information from sensor images, e.g., face identification. See figure 7 of the appendix for a summary of the entire optimization scheme.

3.1 Sensor Layer

Like [57], our sensor layer consists of a conventional imaging system with a fixed focusing lens and learnable phase mask positioned in the aperture plane. Accordingly, we follow [57] and employ computational Fourier optics [17] to model the sensor via a pupil function:

$$P_{\lambda,z}(x_1,y_1) = A(x_1,y_1)\underbrace{e^{-jk_\lambda\left(\frac{x_1^2+y_1^2}{2}\right)\left(\frac{1}{z}-\frac{1}{u}\right)}}_{\phi_{lens}}\left[\underbrace{e^{-jk_\lambda \Delta_n h(x_1,y_1)}}_{\phi_{mask}}+\nu\right] \tag{1}$$

where $A, \phi_{mask}, \phi_{lens}, h \in \mathbb{R}^{W_1 \times H_1}$ denote the amplitude modulation due to the aperture, the phase modulations due to the phase mask and lens, and the learnable heights of the phase mask pixels respectively; z denotes the scene point distance; u the focal plane distance of the lens; f the focal length the lens; $k_\lambda = \frac{2\pi}{\lambda}$ the wave number; and Δ_n the difference between the refractive indices of air and the phase mask material. In an important deviation from [8], we introduce a new variable $\nu > 0$ into the sensor model to account for the portion of light that travels through the phase mask undiffracted. The reason for this is that undiffracted light may, as we show in section 4.1, leak privacy-sensitive information if not accounted for. Since our goal is to design a sensor that optically filters out privacy-sensitive information, keeping track of the undiffracted light is critical. Finally, let $I_\lambda \in \mathbb{R}^{W_2 \times H_2}$ and $M \in \mathbb{R}^{W_2 \times H_2}$ denote an all-in-focus image and its corresponding depth map respectively. Then, the image formed by the sensor layer is

$$I'_\lambda(x_2, y_2) = \sum_{i=1}^{N} \Big[I_\lambda(x_2, y_2) \cdot B_s(\mathbf{1}_{M(x_2,y_2)=z_i}) \Big] * \underbrace{\left| \mathcal{F}\left\{ P_{\lambda,z}\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f} \right) \right\} \right|^2}_{PSF_{\lambda,z}} \tag{2}$$

where $\mathcal{F}$ denotes the discrete Fourier transform; $*$ the convolution operator; $z_1, ..., z_N$ a set of discrete depths; s the size of PSF_{λ,z_i}; $\mathbf{1}_{M(x_2,y_2)=z_i} \in \mathbb{R}^{W_2 \times H_2}$ an indicator function that is true when $M(x_2, y_2) = z_i$; and B_j a max-pool operation with a kernel of size $j \times j$. Note, we normalize over all N depths such that $\sum_{i=1}^{N} B_s(\mathbf{1}_{M(x_2,y_2)=z_i}) = 1$.

Optimization The sensor layer $S : \mathbb{R}^{W_2 \times H_2 \times 3} \to \mathbb{R}^{W_2 \times H_2 \times 3}$ maps an all-in-focus image I to a sensor image $I' = S(I)$. Our goal is to optimize heights of the phase mask $h \in \mathbb{R}^{W_1 \times H_1}$ such that the sensor images I' cannot be used for estimation of sensitive attributes $g(I) \in \mathbf{G}$, but can be used for estimation of the target attributes $t(I) \in \mathbf{T}$. To achieve this, we employ an adversarial training formulation in which the sensor layer is trained to simultaneously promote the success of UtilityNet $U : \mathbb{R}^{W_2 \times H_2 \times 3} \to \mathbf{T}$ while inhibiting the success of AttackNet $A : \mathbb{R}^{W_2 \times H_2 \times 3} \to \mathbf{G}$.

Let L_U and L_A denote the loss functions for UtilityNet and AttackNet respectively. Then, the objective function for the sensor layer is given by

$$L_S(I) = \min_h L_U\big(t(I), U(I')\big) - \eta L_A\big(g(I), A(I')\big), \tag{3}$$

where η denotes a weight parameter to balance the privacy and utility trade off. To implement this loss function, we apply alternating gradient updates to the height map (sensor layer) and the weights of the downstream networks: In step 1, we update the height map and weights of UtilityNet together and in step 2, we update the weights of AttackNet while the sensor layer and Utilitynet are fixed. For our experiments, we set $\eta = 0.01$ and we used the Adam optimizer [26] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon =$ 1e-8 and a learning rate of 0.001. We also bound the heights of the phase mask pixels between [0,1.525]μm by applying a *hardsigmoid* function and then a scaling operation to the height map h.

3.2 Downstream Neural Networks

Downstream of the sensor layer, we have two neural networks, UtilityNet and AttackNet. We define the utility task as monocular depth estimation and the attack task as face identification. Thus, the expected effect of our learned phase mask is to obfuscate identifiable facial information, while boosting the depth estimation accuracy.

For UtilityNet, we adopt the ResNet-based multi-scale network proposed by [60] and we initialize the model with the pre-trained weights from [44]. For optimization of UtilityNet, we follow [1] and adopt an objective function consisting of a weighted sum of losses on the depth, gradient and perceptual quality:

$$L_U(y, \hat{y}) = \frac{1 - \mathrm{SSIM}(y, \hat{y})}{2} + \frac{1}{n}\left[\xi|y - \hat{y}| + |\mathbf{g_x}(y, \hat{y})| + |\mathbf{g_y}(y, \hat{y})|\right] \tag{4}$$

where $y = 10/y_{\mathrm{gt}}$ denotes the reciprocal of the ground-truth depth map, $\hat{y}$ the estimated depth map, and ξ a weighting parameter (which we set to 0.1), n the number of pixels in the depth map, $SSIM$ structural similarity [54] and $\mathbf{g_x}$ and $\mathbf{g_y}$ compute the differences of the x and y components of gradients of y and $\hat{y}$.

For AttackNet, we use the EfficientNet-b0 [50] architecture and adopt a softmax activation followed by a cross-entropy loss for n-way classification as in [5]. For testing, we remove the final layer of the network and learn one-vs-all SVM classifiers for each test subject, using a held-out subset of the evaluation set, as in [5]. Finally, we train both UtilityNet and AttackNet until saturation using the Adam optimizer [26] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon =$ 1e-8 and a learning rate of 0.001.

3.3 Prototype Sensor

The prototyping pipeline accepts a learned phase mask height map and culminates with fine-tuning of UtilityNet and AttackNet with the calibrated PSFs (Figure 8 of appendix). The phase mask is fabricated using two-photon lithography and inserted into the aperture plane of a conventional lens system. Section A of the appendix includes the fabrication details of the prototype for interested readers.

4 Experimental Results

This section includes both simulation and real results with our prototype sensor. For all reported results, attack models are retrained after adversarial optimization is complete, i.e., after the learnable parameters of the sensor layer have been permanently fixed. This ensures that the sensor layer cannot be overcome by an adversary with access to a large set of labeled sensor images. Utility models are also retrained after the sensor layer is fixed. The full details about the datasets, evaluation protocols and image formation model are available in sections B.1, B.2, and B.3 of the appendix respectively.

4.1 Design Space Analysis

We show a systematic analysis of the sensor design space which is essential in designing any optical pre-capture privacy sensor. The utility task is fixed to monocular depth estimation on the NYUv2 dataset [33] and the attack task to face identification on the VGGFace2 dataset [5]. For face identification, the faces were resized, as discussed in B.3, to simulate different camera-to-subject distance between $1 - 10m$ and the face identification performance was averaged over these depth ranges.

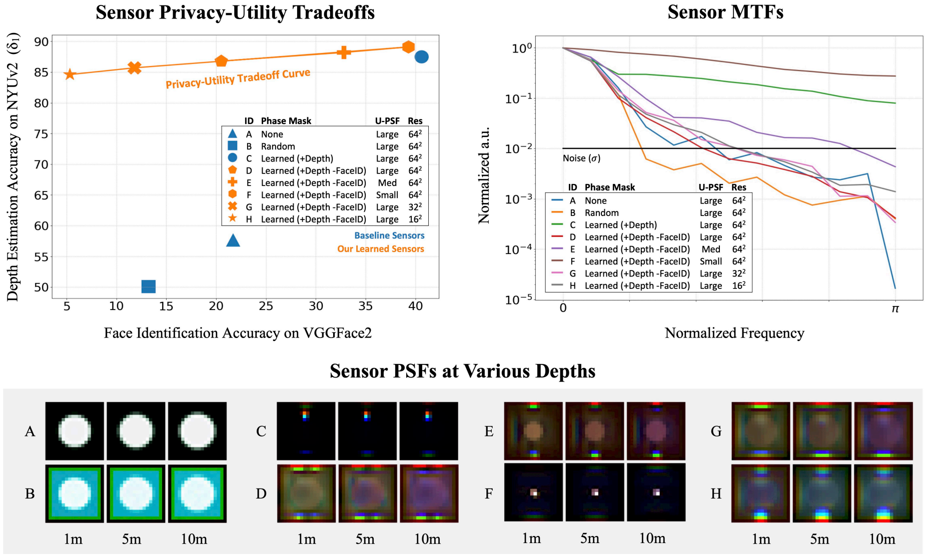


Fig. 2. Design Space Analysis. The plot on the top left shows depth estimation (utility) vs face identification (privacy) accuracy for multiple sensor designs (A-H), with varying phase mask designs, undiffracted PSF sizes and sensor resolutions. The sensors optimized using our approach (D-H), have a better privacy utility trade-off (smaller slope) compared to the traditional sensor designs (A-C). The corresponding modulus-transfer-function (MTF) plots, shown in the top right plot, give an intuition into the privacy performance of the different sensor designs based on their respective cut-off frequencies. Lower cut-off frequency corresponds to better privacy (filtering of facial details). The PSFs of each sensor designs for 3 different depths are shown at the bottom of the figure. Note, the PSFs corresponding to learned sensors vary significantly over different depths, which results in better depth estimation performance.

Impact of Phase Mask Design This section examines how the goal of balancing privacy and utility objectives can be achieved by optimizing the phase mask. We compare four sensors with identical parameters, but with four different phase mask designs. The four phase mask designs that we consider are: (A) none; (B) random; (C) optimized to maximize depth estimation performance; and (D) optimized to maximize depth estimation performance while inhibiting face identification. At the top left of figure 2, we show the privacy-utility tradeoff of each sensor, i.e., the performance of downstream depth estimation and face identification models for each sensor. At the top right of figure 2, we show the modulation-transfer-function (MTFs) corresponding to the PSFs of each sensor at 1 meter. The MTFs are computed as the radially averaged magnitude of the frequency spectrum of the PSFs. The cut-off frequencies are determined by the noise level of $\sigma = 0.01$, which we assume to be Gaussian. At the bottom of figure 2, we show pixel-space visualizations of the depth-dependent point-spread-functions (PSFs) of each sensor. The presence of the large defocus spot in the

PSFs of sensors A, B, C, and D is due to fact that all four sensors are focused at 10cm and a portion of the light travels through the phase masks undiffracted [52]. For all four sensors, focal length f = 16mm, focal distance u = 10cm, sensor size of 8.8×6.6mm, sensor resolution of 64×64 pixels, aperture diameter d = 8.7mm, phase mask pitch 4.25μm, and we considered a working range of $z = [1, 10]$m.

Consider sensor A (no phase mask/naive defocus). Since it is focused at 10cm and does not have a phase mask, faces located in the working range of 1-10m with be heavily out-of-focus. This is advantageous for privacy as high frequency facial details will be optically filtered out. This is consistent with figure 2, as the PSFs of sensor A resemble a large Gaussian filter and the MTF a low-pass filter. Finally, privacy-utility curve in figure 2 shows that sensor A succeeds at reducing face identification accuracy from 80.1% (all-in-focus) images to 21.7%, but fails to provide satisfactory depth estimation performance, as the PSFs remain constant over depth, and thus fails to provide a desirable privacy-utility trade-off.

Consider sensor B (random phase mask). Due to the fact that approximately 10% of the incident light passes through the phase mask undiffracted, its PSF consist of a linear combination of a diffracted and an undiffracted PSF, as shown in figure 2. The undiffracted PSFs exactly match the PSFs of sensor A, which resemble a Gaussian filter. The diffracted PSFs are the result of light passing through the random phase mask, which disperses light uniformly to the entire receptive field, so the diffracted PSF is effectively a square average filter. Comparing the MTFs of sensors A and B in figure 2, we can observe that the cut-off frequency of sensor B is much lower than sensor A. As expected, this results in sensor B having worse downstream depth estimation and face identification performance compared to sensor A. Overall, sensor B fails to provide a desirable privacy-utility trade-off.

Consider sensor C (phase mask optimized to maximize depth estimation performance as in [57]). Although its PSFs also consist of a linear combination of diffracted and undiffracted PSFs, this is obscured in figure 2, by the fact that the diffracted PSFs are very sparse, which results in the "activated" pixels having a much higher magnitude than the undiffracted PSFs. Looking now at the MTF of sensor C in figure 2, we observe that optimizing the phase mask for depth estimation resulted in PSFs that don't filter out any information, which is consistent with what one would expect. Interestingly, we also see from figure 2 that the PSFs vary with depth, which is also what we would expect if our goal is to maximize downstream depth estimation performance. Finally, looking at figure 2, we see that depth estimation performance is comparable to the state-of-the-art, but that it comes at the cost of face identification accuracy also being high (40%). Note, the reason face identification accuracy is not higher than 40% is that sensor C has a resolution of 64 × 64 pixels.

Consider sensor D (learned using our proposed adversarial optimization algorithm to maximize depth estimation performance while minimizing face identification performance). Its PSFs filter out high frequency facial details, yet also vary significantly with depth, as shown in the PSFs and MTF of sensor D in figure 2. Both of these outcomes are intuitively consistent with our goal of balancing privacy and utility. The privacy-utility plot in figure 2 confirms this intuition by showing that downstream depth estimation performance is comparable to the state-of-the-art and that face identification is limited to an accuracy of 20.5%.

Impact of Undiffracted Light vis-à-vis focus settings In practice, ν from equation 1 typically varies usually between 0.08 to 0.2 [52], which results in a non-insignificant amount of undiffracted light reaching the sensor. Previous deep optics works for depth estimation [8, 57] have simply ignored this issue. This was possible because they were not concerned with leakage of privacy sensitive information. For our setting, preventing leakage of privacy sensitive information is crucial, so the undiffracted light must be modeled. We illustrate this by optimizing the phase masks of three sensors (D, E and F) with undiffracted PSFs of different sizes and comparing the resulting privacy-utility trade-offs of the respective sensors. Note, since the undiffracted PSFs don't depend on the phase mask design, they can be fixed prior to optimizing the masks. We vary the size of the undiffracted PSFs by setting the focal distance $u = \{0.1, 0.17, 1.0\}$m to produce sensors with "large", "medium", and "small" undiffracted PSFs. For all three sensors, we fix $\nu = 0.1$, focal length $f = 16$mm, sensor size of 8.8×6.6mm, sensor resolution of 64×64 pixels, aperture diameter $d = 8.7$mm, phase mask pitch 4.25μm, and we considered a working range of $z = [1, 10]$m. The PSFs of each sensor are shown at the bottom of figure 2. In the MTF plot in figure 2, we see that larger undiffracted PSFs result in less leakage of privacy sensitive information. Intuitively, this is reasonable as a larger defocus kernel corresponds to a lower cut-off frequency, so more information will be filtered out, and this is consistent with the results shown in privacy-utility trade-off plot in figure 2. Thus, when designing our final sensor, we utilize a large undiffracted PSF.

Impact of Sensor Resolution Our goal is to design a sensor that inhibits recovery of privacy sensitive information from encoded sensor images. From an attacker's perspective, recovery of sensitive information from encoded images can be modeled as a conventional inverse problem, so the number of observations (or sensor resolution) naturally plays an important role. We study this role by optimizing the phase masks of three sensors with identical parameters, but different resolutions. The three sensor resolutions we consider are 16×16, 32×32 and 64×64, and the resulting privacy-utility trade-offs are shown in the top left of figure 2. As expected, the lower the sensor resolution, the lower the face identification accuracy. Interestingly, the ability of the downstream depth estimation models to produce high quality depths maps of size 256×256 is not meaningfully impacted by the sensor resolution as scene depth in most natural settings tends to be a low frequency signal. This is highly advantageous for our setting as we are able to reduce face identification performance by reducing the sensor resolution without sacrificing significant depth estimation performance. Note, for ease of comparison, we display all the sensor PSFs and MTFs using a

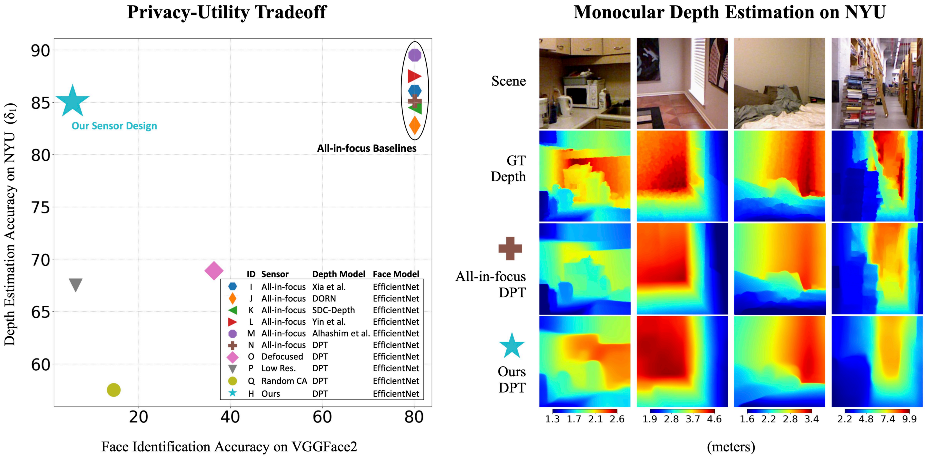


Fig. 3. Comparisons with Existing Methods. On the left, we compare the privacy-utility trade-off of our learned sensor design against four other sensors: a conventional all-in-focus sensor, a defocused privacy sensor [40], a low resolution privacy sensor [11] and a random coded aperture privacy sensor [55]. For all sensors, we show the corresponding face identification accuracy of EfficientNet-b0 [50] on the VGGFace2 dataset, along the x-axis, and the depth estimation accuracy (δ_1) of DPT [43] on the NYU Depth v2 dataset, along the y-axis. For the all-in-focus sensor, we additionally show the depth estimation performance of [59, 16, 53, 64, 1]. On the right, we compare the monocular depth estimation predictions of DPT operating on conventional all-in-focus images vs private images from our learned sensor design.

sensor resolution of 64×64. For all three sensors, focal length $f = 16$mm, sensor size of 8.8×6.6mm, aperture diameter $d = 8.7$mm, phase mask pitch 4.25μm, and we considered a working range of $z = [1, 10]$m.

4.2 Simulation Results

Comparisons with Pre-Capture Privacy Sensors We compare the privacy-utility trade-off provided by our optimized sensor design (sensor H from figure 2) to an all-in-focus sensor and three existing pre-capture privacy sensors: a heavily defocused sensor [40], an extremely low-resolution sensor [11], and a sensor with a random coded aperture mask [55]. For all four sensors, the focal length $f = 16$mm, the sensor size was 8.8×6.6mm, and a working range of $z = [1, 10]$m was considered. For the all-in-focus sensor (N), the focal distance $u = \infty$, the sensor resolution was 256×256 pixels, and the aperture diameter $d = 1$mm. For the defocused sensor (O), the focal distance $u = 10$cm, the sensor resolution was 256×256 pixels, and the aperture diameter $d = 8.7$mm. For the low resolution sensor (P), the focal distance $u = \infty$, the sensor resolution was 16×16 pixels, and the aperture diameter $d = 1$mm. Lastly, for the coded aperture camera (Q), the focal distance $u = 10$cm, the sensor resolution was 256×256 pixels, and the aperture diameter $d = 8.7$mm. For fair evaluation, new

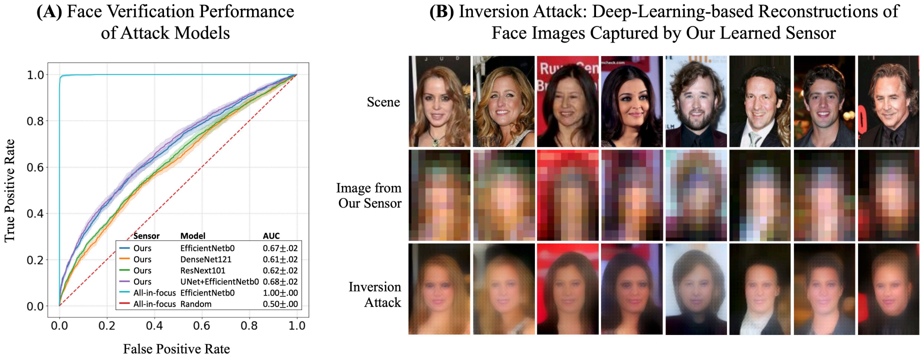


Fig. 4. Privacy Evaluation. Figure **(A)** shows the face verification performance of different attack models on our sensor simulated LFW dataset. A random classifier and EfficientNet-b0 [50] trained on an all-in-focus LFW dataset serve as our lower and upper bounds respectively. For figure **(B)**, we train a U-Net [29] style neural network to reconstruct the original source images from simulated sensor images. The results show that the U-Net learns to reconstruct low frequency information such as hair color and skin pigment, but fails to reconstruct the same identity as the original person, as the high frequency identity information is filtered out by our optimized sensor.

copies of UtilityNet and AttackNet were trained from scratch to test each sensor, using only images produce by the respective sensor. As shown on the left of figure 3, our data-driven approach (H), achieves a far better privacy-utility trade-off compared to the three pre-capture sensors (O, P and Q) and the all-in-focus baseline (N). It is important to note that the sensor resolution and defocus settings of our learned sensor (H) are the same as the low-res (P) and defocused (O) sensors respectively. This clearly demonstrates the advantage of our learning based approach over the fixed sensors.

For completeness, we also compare our learned sensor design against six state-of-the-art monocular depth estimation methods that operate on all-in-focus images from conventional sensors. The six methods we compare against are Xie et al. [59], DORN [16], SDC-depth [53], Yin et al. [64], Alhashim et al. [1] and DPT [43]. As illustrated in figure 3, our approach limits face identification performance to 5.3% compared to 80.1% for the all-in-focus sensor, while still achieving depth estimation performance comparable to the state-of-the-art. On the right side of figure 3, we qualitatively compare the predicted depth maps of DPT [43] when operating on all-in-focus sensor images vs private images from our learned sensor. Here we note the limitation of our approach as the depth maps produced by our approach doesn't preserve very high frequency details.

Privacy Evaluation In this section, we assess the efficacy of our learned sensor design (H) at inhibiting facial identification attacks by four different models: EfficientNetb0 [50], DenseNet121 [21], ResNext101 [62], and UNet+EfficientNetb0. All models are retrained on the private images from our sensor. In the fourth

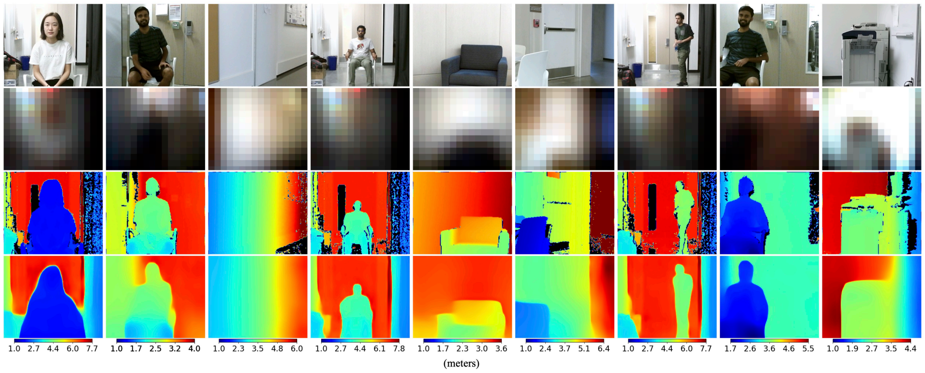


Fig. 5. Private Depth Estimation with Prototype Sensor. We show the depth reconstruction performance of our prototype sensor in the wild. Row 1: all-in-focus images of the scene. Row 2: Images captured by our prototype sensor (16×16 pixels). Row 3: Depth map captured by Microsoft Kinect v2. Row 4: Depth maps predicted from images captured by our prototype sensor. The mean depth estimation accuracy (δ_1) of our predicted depth maps is 83.73%. This is consistent with the 84.69% accuracy of our simulated results on the NYUv2 Dataset.

model (UNet+EfficientNetb0), the UNet [29] precedes EfficientNetb0 and is trained to reconstruct the original face images from their encoded counterparts (i.e., from simulated captures from our learned sensor). Figure 4(A) shows the performance of the four models in the form of receiver operating characteristic (ROC) curves and figure 4(B) shows some sample reconstructions from the UNet. Our learned sensor limits face verification performance to an area-under-the-curve (AUC) of 0.67 ± 0.02 for the best performing model. This represents a significant obfuscation of face identity information using our sensor design, considering that the same network, when learned on images from a conventional sensor, achieves an AUC of 0.99 ± 0.05. Regarding the reconstructions shown in figure 4(B), we can see that while it's possible to recover some low frequency information, such hair color and skin pigment, from a sensor image, key high frequency features, such as the lips, eyes and nose, are incorrectly reconstructed, which prevents successful facial identification. The full details of the evaluation are provided in section B.4 of the appendix.

Privacy-Preserving Action Recognition We further evaluate the utility of our learned sensor design (H) by training a 3D-fused two-stream model (I3D) [6] for action recognition using simulated color images and predicted depth maps from our learned sensor. For comparison, we also train an I3D model on conventional all-in-focus color images and "ground-truth" depth maps from a Microsoft Kinect v2. Both models are trained on the NTURGBD120 [28] dataset using the cross-setup training and testing protocol. The model trained on outputs from the Kinect achieved a top 1 and top 5 accuracy of 79.1% and 94.0% respectively. The model trained on outputs from our learned sensor achieved a top 1

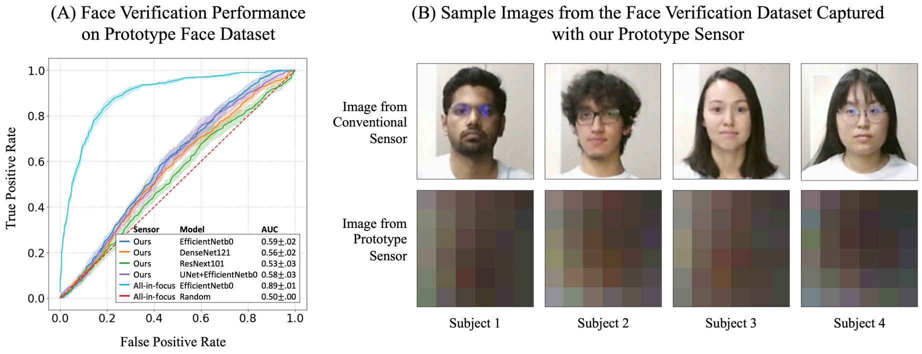


Fig. 6. Privacy Evaluation of Prototype Sensor. (A) Face verification performance of different attack models on the face images from a dataset of 20 individuals captured by our sensor prototype. A random classifier and the EfficientNet-b0 attack on all-in-focus images serve as our lower-bound and upper-bound respectively. **(B)** Sample Images from the face verification dataset captured with our prototype sensor (row 2) with corresponding all-in-focus images (row 1).

and top 5 accuracy of 70.5% and 91.5% respectively. These results demonstrate that out optimized sensor has the potential impact a range of applications for which privacy is a major concern, such as elder care, remote health monitoring, surveillance of sensitive environments (schools, hospitals, etc.), and more.

4.3 Results with Real Prototype Sensor

To demonstrate the viability of our approach, we build a physical prototype (shown in figure 1) of our optimized sensor design (sensor H from figure 2) and evaluate it's performance along a range of dimensions. In order to avoid fabricating a low resolution sensor we simply downsample a high resolution sensor to 16 × 16 pixels. The consequences of such a choice are discussed in detail along with the prototyping pipeline in section A of the appendix.

Privacy-Preserving Depth Estimation with Prototype Sensor We present qualitative depth estimation results on real captures from our prototype sensor in row 4 of figure 5. The mean depth estimation accuracy (δ_1) for these results is 83.73%, which is consistent with the 84.69% accuracy of our simulated results on the NYUv2 Dataset. Images of the corresponding scene captured with a Kinect color camera and our prototype sensor (16 × 16 pixels) are shown in rows 1 and 2 respectively, and depth images captured from a stereo calibrated time-of-flight Kinect sensor are shown in row 3. Qualitatively, the depth estimates from our sensor are comparable to the Kinect measurements, but lack some of the high frequency details.

Privacy Evaluation of Prototype Sensor To validate our prototype sensor's ability to inhibit face identification, we capture a novel face verification dataset using our prototype sensor, consisting of 100 images of 20 subjects (5 images per subject at different depths). As an upper bound, we also capture an identical dataset using a conventional all-in-focus color camera. For the evaluation, we generate 10 sets of 200 pairs of face images for 10-fold cross-validation. Sample images from our dataset are shown in figure 6(B).

For the evaluation, we assume a white-box attack model and use the same protocol as for our previous simulation-based privacy analysis. The results of the evaluation, presented in figure 6(A), show that our prototype sensor limits face verification performance to an area-under-the-curve (AUC) of 0.59 ± 0.02 for the best performing model. The same model, when operating on images from a conventional sensor, achieves an AUC of 0.89 ± 0.01. These results demonstrate that we are able to reproduce our simulated results with a real hardware prototype.

5 Conclusions

We believe our framework and prototype sensor design represent a first and significant advance towards enabling a new generation of pre-capture privacy aware computational cameras that will greatly expand the range of environments, technologies, and applications where computer-vision-based solutions can be deployed. Thus, it becomes important to discuss what other possible utility and privacy tasks our proposed framework can be applied to, to get good privacy-utility trade-offs. Based on our analysis of the MTFs of various sensor designs (figure 2), a good pair of privacy-utility tasks would be one with contrasting requirements for frequency/detail preservation. For example, for a privacy task of inhibiting face identification, a utility task of object classification is likely to work best for objects that are larger than a human face, such as a human body, furniture, cars, etc.

Our choice of utility and privacy objectives is a rather interesting one. The optimization makes sure that the psfs produced vary over depth, but still act as a low pass filter. This enables us to estimate high quality depth maps, while inhibiting face identification. However, it's also important to note a limitation of our choice of privacy-utility tasks. Namely, that the objective of acquiring high-frequency depth maps comes into direct conflict with the objective of preventing accurate face identification, resulting in over smoothed depth estimates. Nevertheless, as shown in section 4.2, our estimated depth maps have enough detail for many downstream vision tasks such as depth-based activity recognition.

References

1. Alhashim, I., Wonka, P.: High quality monocular depth estimation via transfer learning. arXiv preprint arXiv:1812.11941 (2018)
2. Alvi, M., Zisserman, A., Nellåker, C.: Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. Proceedings of the European Conference on Computer Vision (ECCV) pp. 0–0 (2018)
3. Beach, S., Schulz, R., Downs, J., Matthews, J., Barron, B., Seelman, K.: Disability, age, and informational privacy attitudes in quality of life technology applications: Results from a national web survey. ACM Transactions on Accessible Computing (TACCESS) **2**(1), 5 (2009)
4. Boominathan, V., Adams, J.K., Robinson, J.T., Veeraraghavan, A.: Phlatcam: Designed phase-mask based thin lensless camera. IEEE transactions on pattern analysis and machine intelligence **42**(7), 1618–1629 (2020)
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) pp. 67–74 (2018)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
7. Chakrabarti, A.: Learning sensor multiplexing design through back-propagation. Advances in Neural Information Processing Systems pp. 3081–3089 (2016)
8. Chang, J., Wetzstein, G.: Deep optics for monocular depth estimation and 3d object detection. Proceedings of the IEEE International Conference on Computer Vision pp. 10193–10202 (2019)
9. Chen, J., Konrad, J., Ishwar, P.: Vgan-based image representation learning for privacy-preserving facial expression recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops pp. 1570–1579 (2018)
10. Chhabra, S., Singh, R., Vatsa, M., Gupta, G.: Anonymizing k-facial attributes via adversarial perturbations. arXiv preprint arXiv:1805.09380 (2018)
11. Dai, J., Wu, J., Saghafi, B., Konrad, J., Ishwar, P.: Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops pp. 68–76 (2015)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on pp. 248–255 (2009)
13. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. Advances in Neural Information Processing Systems pp. 658–666 (2016)
14. Dwork, C.: Differential privacy: A survey of results. International conference on theory and applications of models of computation pp. 1–19 (2008)
15. Erdélyi, A., Barát, T., Valet, P., Winkler, T., Rinner, B.: Adaptive cartooning for privacy protection in camera networks. 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) pp. 44–49 (2014)
16. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 2002–2011 (2018)

17. Goodman, J.W.: Introduction to Fourier optics. Roberts and Company Publishers (2005)
18. Haim, H., Elmalem, S., Giryes, R., Bronstein, A.M., Marom, E.: Depth estimation from a single image using deep learned phase coded mask. IEEE Transactions on Computational Imaging **4**(3), 298–310 (2018)
19. He, L., Wang, G., Hu, Z.: Learning depth from single images with deep neural network embedding focal length. IEEE Transactions on Image Processing **27**(9), 4676–4689 (2018)
20. Hinojosa, C., Niebles, J.C., Arguello, H.: Learning privacy-preserving optics for human pose estimation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) pp. 2573–2582 (October 2021)
21. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
22. 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)
23. Jeong, Y., Yoo, D.H., Cho, J., Lee, B.: Optic design and image processing considering angle of incidence via end-to-end optimization method. Ultra-High-Definition Imaging Systems II **10943**, 109430U (2019)
24. Jia, S., Lansdall-Welfare, T., Cristianini, N.: Right for the right reason: Training agnostic networks. International Symposium on Intelligent Data Analysis pp. 164–174 (2018)
25. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: Training deep neural networks with biased data. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 9012–9020 (2019)
26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
27. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. CVPR **2**(3), 4 (2017)
28. Liu, J., Shahroudy, A., Perez, M.L., Wang, G., Duan, L.Y., Chichung, A.K.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence (2019)
29. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
30. Metzler, C.A., Ikoma, H., Peng, Y., Wetzstein, G.: Deep optics for single-shot high-dynamic-range imaging. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 1375–1385 (2020)
31. Mirjalili, V., Raschka, S., Ross, A.: Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers. 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS) pp. 1–10 (2018)
32. Mirjalili, V., Raschka, S., Ross, A.: Flowsan: privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers. IEEE Access **7**, 99735–99745 (2019)
33. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. ECCV (2012)

34. Nawaz, T., Rinner, B., Ferryman, J.: User-centric, embedded vision-based human monitoring: A concept and a healthcare use case. Proceedings of the 10th International Conference on Distributed Smart Camera pp. 25–30 (2016)
35. Neustaedter, C.G., Greenberg, S.: Balancing privacy and awareness in home media spaces. Citeseer (2003)
36. Nguyen Canh, T., Nagahara, H.: Deep compressive sensing for visual privacy protection in flatcam imaging. Proceedings of the IEEE International Conference on Computer Vision Workshops pp. 0–0 (2019)
37. 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)
38. Phan, B., Mannan, F., Heide, F.: Adversarial imaging pipelines. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16051–16061 (2021)
39. Pittaluga, F., Koppal, S., Chakrabarti, A.: Learning privacy preserving encodings through adversarial training. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) pp. 791–799 (2019)
40. Pittaluga, F., Koppal, S.J.: Privacy preserving optics for miniature vision sensors. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 314–324 (2015)
41. Pittaluga, F., Koppal, S.J.: Pre-capture privacy for small vision sensors. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2215–2226 (2016)
42. Pittaluga, F., Zivkovic, A., Koppal, S.J.: Sensor-level privacy for thermal cameras. 2016 IEEE International Conference on Computational Photography (ICCP) pp. 1–12 (2016)
43. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ArXiv preprint (2021)
44. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)
45. Sattar, H., Krombholz, K., Pons-Moll, G., Fritz, M.: Shape evasion: Preventing body shape inference of multi-stage approaches. arXiv preprint arXiv:1905.11503 (2019)
46. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
47. 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)
48. Sun, Q., Tseng, E., Fu, Q., Heidrich, W., Heide, F.: Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1386–1396 (2020)
49. Tan, J., Khan, S.S., Boominathan, V., Byrne, J., Baraniuk, R., Mitra, K., Veeraraghavan, A.: Canopic: Pre-digital privacy-enhancing encodings for computer vision. 2020 IEEE International Conference on Multimedia and Expo (ICME) pp. 1–6 (2020)
50. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning pp. 6105–6114 (2019)
51. Tseng, E., Mosleh, A., Mannan, F., St-Arnaud, K., Sharma, A., Peng, Y., Braun, A., Nowrouzezahrai, D., Lalonde, J.F., Heide, F.: Differentiable compound optics and processing pipeline optimization for end-to-end camera design. ACM Transactions on Graphics (TOG) **40**(2), 1–19 (2021)

52. Wang, H., Liu, Y., Ruan, Q., Liu, H., Ng, R.J., Tan, Y.S., Wang, H., Li, Y., Qiu, C.W., Yang, J.K.: Off-axis holography with uniform illumination via 3d printed diffractive optical elements. Advanced Optical Materials **7**(12), 1900068 (2019)
53. Wang, L., Zhang, J., Wang, O., Lin, Z., Lu, H.: Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 541–550 (2020)
54. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
55. Wang, Z.W., Vineet, V., Pittaluga, F., Sinha, S.N., Cossairt, O., Bing Kang, S.: Privacy-preserving action recognition using coded aperture videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops pp. 0–0 (2019)
56. Winkler, T., Erdélyi, A., Rinner, B.: Trusteye. m4: protecting the sensor—not the camera. 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) pp. 159–164 (2014)
57. Wu, Y., Boominathan, V., Chen, H., Sankaranarayanan, A., Veeraraghavan, A.: Phasecam3d—learning phase masks for passive single view depth estimation. In: 2019 IEEE International Conference on Computational Photography (ICCP). pp. 1–12. IEEE (2019)
58. Wu, Y., Yang, F., Ling, H.: Privacy-protective-gan for face de-identification. arXiv preprint arXiv:1806.08906 (2018)
59. Xia, Z., Sullivan, P., Chakrabarti, A.: Generating and exploiting probabilistic monocular depth estimates. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 65–74 (2020)
60. Xian, K., Shen, C., Cao, Z., Lu, H., Xiao, Y., Li, R., Luo, Z.: Monocular relative depth perception with web stereo data supervision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 311–320 (2018)
61. Xiao, T., Tsai, Y.H., Sohn, K., Chandraker, M., Yang, M.H.: Adversarial learning of privacy-preserving and task-oriented representations. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)
62. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
63. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST) **10**(2), 1–19 (2019)
64. Yin, W., Liu, Y., Shen, C., Yan, Y.: Enforcing geometric constraints of virtual normal for depth prediction. Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 5684–5693 (2019)
65. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
66. Zhuang, Z., Bradtmiller, B.: Head-and-face anthropometric survey of us respirator users. Journal of occupational and environmental hygiene **2**(11), 567–576 (2005)