Physics Based Camera Privacy: Lens and Network Co-Design to the Rescue

Marius Dufraisse
DTIS-ONERA, France
marius.dufraisse@onera.fr

Marcela Carvalho
Upciti, France
marcela.carvalho@upciti.com

Pauline Trouvé-Peloux
DTIS-ONERA, France
pauline.trouve@onera.fr

Frédéric Champagnat
DTIS-ONERA, France
frederic.champagnat@onera.fr

Abstract

The wide presence of cameras using automatic image processing has a positive impact on smart cities and autonomous driving objectives but also poses privacy threats. In this context, there has been an increasing interest in regulating the acquisition and use of public and private data. In this paper, we propose a method to co-design a lens and an image processing pipeline to perform semantic segmentation of urban images, while ensuring privacy by introducing optical aberrations directly on the lens. This particular design relies on a hardware level of protection that prevents an attacker from accessing sensitive information from the original images of a device. Related works rely on specific privacy threats, requiring a large database for training. In contrast, we propose to seek robustness by preventing deblurring during training in a self-supervised way, thus, without requiring additional annotations. Moreover, we validate our approach by simulating attacks with deblurring, and license plate detection and recognition to show that our model can fool these tasks with success while keeping a high score on the utility task.

1. Introduction

Deep neural networks can perform many computer vision tasks and are likely to be embedded in more and more imaging systems. However, citizens and regulators are aware of the privacy risks posed by the increasing number of connected cameras. To mitigate this issue, processing can be applied to relevant features extracted from the image to allow a utility task (*e.g.*, object detection, instance segmentation) while preventing the reconstruction of sensitive information. Most common approaches are performed at software-level [14, 29], although that does not entirely protect against an attacker who would get access to the sensor data (see Figure 1). To increase the system's safety, it is preferable to use a hardware approach, such as a specific lens to physically blur the image before it reaches the sensor [9, 23].

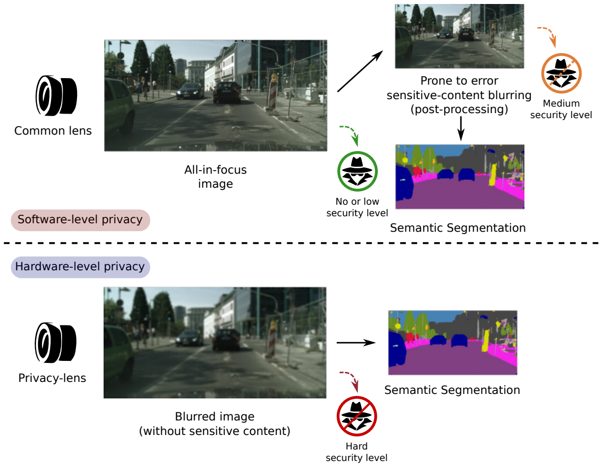

Figure 1. Two approaches to protect privacy. Software-level privacy uses an algorithmic component to mask sensitive data in the image: it fails to ensure privacy if one gets access to the device, or if the sensitive content is not correctly detected. Hardware-level protection guarantees that even if one were to get full access to the camera software, sensitive information would not be extracted.

To go further, the processing and privacy constraints have to be considered when designing the lens. There must be a balance between privacy and processing performance. Thus, if the privacy features are removed, the main task should still be able to be performed.

In this context, camera design can be conducted with optics/neural network co-design methods, *a.k.a* deep optics [10], that jointly optimize lens parameters and neural network weights. This contrasts with the standard approach with fixed optical parameters for sensing and tuning the image processing algorithm for some tasks. In deep optics a parameterized optics encoder is plugged at the input of the task-oriented neural network and parameters of both are optimized thanks to gradient descent and examples from a dataset. A typical optical encoder is depicted in Figure 3. It specifically computes two features with dominant effect on image quality: blur – through the point spread function (PSF) – and noise distribution parameters. Examples from the dataset are convolved with the PSF and noise is added according to the noise model, generating physically sound images. Finally, the simulated image can be fed into a neural network (see Figure 2).

The described pipeline must be differentiable to explore co-design and optimize the optical parameters using gradient descent. In particular, it is necessary to compute not only the PSF but also its gradients with respect to the optical parameters. This can be performed with two main types of simulations: differentiable ray tracing (DRT) and Fourier optics.

Differentiable ray tracing can optimize a lens made of multiple glass elements for Extension of Depth of Field (EDOF) [22, 31, 40] or object detection [13]. Differentiable ray tracing is well suited for the optimization of a set of conventional lenses, but it is a geometric model that does not take diffraction into account.

On the other hand, Fourier optics [19] is adapted to the simulation of diffractive elements. As such, it has been successfully used to optimize a phase mask for EDOF [39], or depth estimation [21]. Fourier Optics has already been adopted to optimize optics for Human Pose Estimation (HPE) while protecting privacy [24]. In this work, a diffractive element is optimized so that the sensor image contains enough information to perform pose estimation while preventing face recognition. Other works have shown that this method can also be applied to scene captioning [6]. However, these approaches use co-design to counter specific privacy threats and require a large database for training on both main and adversarial tasks.

In this paper, we present a co-design method for autonomous driving with privacy-preserving data. Here are our contributions:

- We propose a privacy-lens design framework that can add a robust level of visual privacy protection to existing recognition systems without significant impact on associated non-privacy tasks (*e.g.*, semantic segmentation);
- We adopt an adversarial training strategy to optimize optical parameters, like [23], but we directly use a deblurring network in a self-supervised fashion to prevent our model against adversarial attacks;
- We propose a solution to automatically balance the influence of the utility task and the adversarial examples during training using gradient penalty [20];
- We assess the protection of the proposed encoded images against attack with state-of-the-art networks for deblurring;
- Finally, we validate our privacy-preserving design method on a sensitive task such as license plate detection and recognition (LPDR).

To the best of our knowledge, this is the first work that proposes using deep optics for privacy protection in the scope of urban scenes, and more specifically, autonomous driving.

2. Related Work

Related works are usually classified into two main categories for privacy issues: software and hardware-level protection.

Software-level privacy methods modify an input image after the image acquisition. Usually, they consist of one network that detects sensitive information (e.g., faces, license plates, skin color) and another model that removes or encodes this information by blurring [1], pixelization [17], masking, or fake image generation [11, 14, 25, 26]. Lately, these last techniques have been of great success, as they propose to modify the image while guaranteeing a small impact on the main task. However, these systems depend on the target domain of the training dataset and are prone to error. For example, a person's identity can easily be compromised if the first network fails to detect one's face. Also, a possible attacker can have access to the images from the sensor before anonymization.

Thus, *hardware-level* solutions are usually more robust as they propose to remove privacy features during image acquisition which would make it impossible for an attacker to have the real information. This includes using low-resolution sensors [8, 15], or event cameras [2, 3, 27]. Recent works also propose to design specific privacy-lens [9, 23, 24] to physically filter sensitive information. In [9], the authors explore lens designs and depth information to produce more blur near the sensor than far from it, as objects already lose information with the distance. However, the choice of optical parameters is performed empirically. [23, 24] propose a deep optics approach where the optical encoder is jointly optimized with a neural network decoder for human pose estimation (HPE). More specifically, the camera lens is optimized to add aberrations and, thus, protect the private attributes of a scene. At the same time, a deep network decoder is optimized to extract the HPE from these corrupted images. This project evolves to [23], where the model is remodeled to use an adversarial component during training to prevent post-training attacks, and also temporal similarity matrices (TSM) for temporal consistency between frames. Despite the proposed automation of optical design, the model depends on annotations for the adversarial task, which are based on the classification of private attributes.

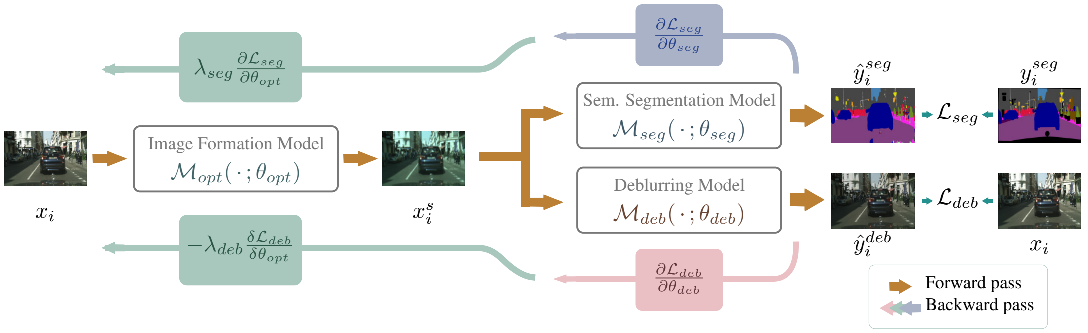
$\lambda_{seg}\frac{\partial \mathcal{L}_{seg}}{\partial \theta_{opt}}$
$\frac{\partial \mathcal{L}_{seg}}{\partial \theta_{seg}}$
Image Formation Model
$\mathcal{M}_{opt}(\cdot;\theta_{opt})$
Sem. Segmentation Model
$\mathcal{M}_{seg}(\cdot;\theta_{seg})$
Deblurring Model
$\mathcal{M}_{deb}(\cdot;\theta_{deb})$
x_i
x_i^s
$\hat{y}_i^{seg}$
$\mathcal{L}_{seg}$
y_i^{seg}
$\hat{y}_i^{deb}$
$\mathcal{L}_{deb}$
x_i
$-\lambda_{deb}\frac{\delta \mathcal{L}_{deb}}{\delta \theta_{opt}}$
$\frac{\partial \mathcal{L}_{deb}}{\partial \theta_{deb}}$
Forward pass
Backward pass

Figure 2. Overview of the proposed adversarial training strategy for camera co-design. In the forward pass, the image formation model simulates the effects of the lens and the sensor on the scene, as described in Section 3.1. Subsequently, each task is performed by a neural network for semantic segmentation (utility task) and deblurring (adversarial task). We detail training of the two antagonist tasks in Section 3.2. In the backward pass (blue and red arrows), one optimization step is applied to update θ_{seg} and θ_{deb}; then, a second pass (green arrow) is used to apply an optimization step to the optical parameters, θ_{opt}.

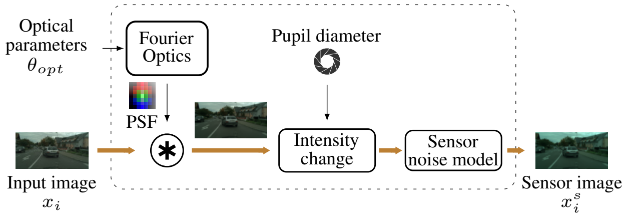

Figure 3. The optical encoder simulates the effect of the optical parameters θ_{opt} through convolution with the PSF. The sensor model introduces an intensity change linked to the aperture and noises.

In this work, we also adopt a co-design training pipeline to select the best optical parameters automatically. However, our adversarial task relies on a deblurring network. This guarantees a self-supervised alternative that has the advantage of avoiding the need for further annotations. Indeed, we rely on the input data (sharp-colored images) for the ground truth. Thus, our approach is more generic as the lens optimization is not dedicated to protecting the image from a specific attack. Also, we provide an ablation study on different techniques from multi-task learning and adversarial training stabilization to find the best trade-off between the training objectives.

3. Proposed Method

In this paper, we propose to co-design a privacy-preserving camera dedicated to urban scene analysis. Specifically, the required task is semantic segmentation, from which further scene analysis can be conducted such as counting cars or pedestrians. We adopt a differentiable optical model based on Fourier optics to jointly optimize the optical parameters and a semantic segmentation network. Also, like in [23,24], we use an adversarial training strategy to introduce privacy knowledge to the optical model. However, instead of relying on private-attributes annotations, we use a deblurring neural network to restore the original inputs and simulate an attack that would start with the deblurring of the generated privacy images from the sensor. We can observe the pipeline of our training strategy in Figure 2. The objective is to generate degraded images which cannot be restored by this deblurring network. Thus, optical parameters should be specialized to predict semantic segmentation maps from the degraded images while the privacy attributes cannot be restored. The proposed pipeline enables using the gradients from the adversarial task to improve the optical encoder parameters, so the images cannot be reconstructed with the original sensitive content.

In the following, we explain our method in further detail, and we also describe our experiments.

3.1. Optical encoder

Image Formation Modeling. We consider an optical system composed of a lens of focal f and a phase mask. The last is a glass element of varying thickness which introduces a phase difference to the incoming light, as illustrated in Figure 4. The phase mask thickness h is parameterized in the Zernike basis of polynomials as commonly used in the literature to describe optical systems [43]. We only consider a sum of the $n = 6$ first Zernike polynomials.

We model the effect of the optics on the scene by com-

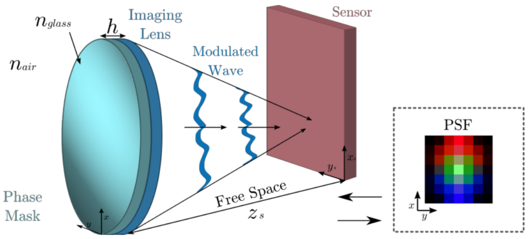

Figure 4. Fourier optics description of the optical system: a planar wavefront reaches the phase mask which introduces a phase delay to the modulated wavefront.

puting its point spread function, which can be convolved with an image, to simulate the effect of the optical part. For the sake of computational complexity, we use Fourier optics [19] to compute the point spread function of the phase mask.

With light sources of wavelength λ far from the aperture, we can consider that the incident wavefront is made of plane waves. In this case, its phase at a point (x, y) of the aperture is $\psi\left(x^2+y^2\right)$ where $\psi = \frac{\pi R^2}{\lambda}\left(\frac{1}{z_o}+\frac{1}{z_s}-\frac{1}{f}\right)$ is the defocus introduced by the lens (with R the aperture radius, z_o and z_s the positions of the object and the sensor with respect to the lens). Then the phase mask of refractive index n_{glass} introduces a phase difference $\phi\left(x, y\right) = \left(n_{glass} - n_{air}\right) h\left(x, y\right)$ compared to propagation in the air (of refractive index n_{air}). As such the complex light amplitude field after the aperture is

$$U_{out}(x, y) = \mathcal{A}\left(x, y\right) e^{i\phi(x,y)} e^{i\psi\left(x^2+y^2\right)} \tag{1}$$

where $\mathcal{A}\left(x, y\right)$ is the transmittance of the aperture and is equal to 1 where it is opened and 0 everywhere else. The propagation between the phase mask and the sensor can then be computed using the Fourier transforms $\mathcal{F}$ to get the field on a point (x_s, y_s) of the sensor. The intensity of the point spread function is obtained by taking the square of the magnitude of this field

$$PSF(x_s, y_s) = \left|\mathcal{F}\left(U_{out}\right)\left(\frac{x_s}{\lambda z_s}, \frac{y_s}{\lambda z_s}\right)\right|^2. \tag{2}$$

This model is differentiable and its implementation in Pytorch enables the joint optimization of optical and neural network parameters using gradient descent.

Sensor model. The sensor is modeled in two steps, as introduced in [16]: an intensity change and a sensor model (Figure 3). The intensity change models the fact that reducing the aperture will lead to a darker image. The sensor model adds non-linearities and multiple sources of noise, in particular Schottky noise which is signal dependent. It also introduces different quantum efficiencies for each channel in the sensor and applies the Bayer filter needed to get 3 channels on a camera. In addition, we apply a simple bilinear demosaicking filter to process RGB images instead of RAW images.

3.2. Optimization Framework

Models. We use SegFormer [42] as semantic segmentation model and we adopt RED-Net [35] for our deblurring objective. SegFormer comprises an encoder-decoder structure with a multi-scale hierarchical Transformer structure for the first part (encoder) and a lightweight multilayer perceptron (MLP), for the second part (decoder). In the original paper, the authors propose a series of Mix Transformer encoders (MiT) that differ in size. We choose to use MiT-B0 which is the smallest as our training pipeline contains 3 models, and is thus already memory-demanding.

RED-Net is a small encoder-decoder image restoration network. We adopt the 10-layer version, referred to as RED10 in the original paper, which does not contain skip connections between the convolution and the deconvolution parts. The encoder is made of four blocks that each divide image resolution by two. The decoder uses transposed convolutions to retrieve the original image resolution. RED-Net has a simpler architecture compared to modern image restoration networks but is well suited for co-design as its smaller receptive field can still invert the local blurring from the PSF and it trains faster.

Adversarial Training. Adversarial training consists of building defenses on a network against attacks by training a model using adversarial examples [37]. In [18], Goodfellow *et al.* proposed a generative adversarial network (GAN), which made this training strategy widely popular. In the paper, two networks were trained simultaneously in a min-max game: a generator network (G) should fool a second network, the discriminator (D), by generating images close to a dataset distribution, while D was trained to classify images from the dataset as real and from G, as fake. One evolution of this strategy is to train directly the neural networks to prevent adversarial attacks that could jeopardize their primary task [4, 33]. Here, we follow this strategy to optimize the parameterized optical model explained in Section 3.1, to concurrently produce images that can be used by a semantic segmentation network, while hiding private attributes. We achieve this property by using a deblurring network to generate adversarial examples and help our optical model become stronger against attacks.

The proposed adversarial training is illustrated in Figure 2, and is explained as follows. In the forward pass, the original image, x_i is modified in the optical encoder, $\mathcal{M}_{opt}$. Then, x_i^s is fed to the semantic segmentation model, $\mathcal{M}_{seg}$, and the deblurring model, $\mathcal{M}_{deb}$. Here, θ_t, where $t \in \{opt, seg, deb\}$ corresponds to each model's parameters. Finally, we calculate the pixel-wise cross-entropy loss, $\mathcal{L}_{seg}$, for semantic segmentation; and we use the mean-squared error (MSE) $\mathcal{L}_{deb}$, for deblurring. The ground truth for the deblurring model corresponds to the original image, x_i before being modified by the optical model. In the next step, both θ_{seg} and θ_{deb} are optimized with their respective gradients $\partial\mathcal{L}_{seg}/\partial\theta_{seg}$ and $\partial\mathcal{L}_{deb}/\partial\theta_{deb}$. And finally, we update the optical parameters by trying to minimize $\mathcal{L}_{seg}$ and concurrently maximize $\mathcal{L}_{deb}$, with respect to θ_{opt}. This translates to:

$$\min_{\theta_{opt}}(\lambda_{seg}\mathcal{L}_{seg}(\hat{y}_i^{seg}, y_i^{seg}) - \lambda_{deb}\mathcal{L}_{deb}(\hat{y}_i^{deb}, x_i)) \tag{3}$$

let $\hat{y}_i^{seg} = \mathcal{M}_{seg}(x_i^s; \theta_{opt}, \theta_{seg})$, and correspondingly $\hat{y}_i^{deb} = \mathcal{M}_{deb}(x_i^s; \theta_{opt}, \theta_{deb})$ be the outputs of the semantic segmentation and deblurring networks. y_i^{seg} is the groundtruth from the semantic segmentation dataset. Also, $\lambda_t \geq 0$ are the weights used to balance each task influence on the optical model optimization. All our experiments consider that we have no specific loss or regularization for the optical model, thus, there is no λ_{opt}. Indeed, by using two different models to optimize a third one, we can relate to a multi-objective optimization (MOO) problem. This means we need to find the best way to counterbalance their effects on the common parameters and reach the best results.

This can be achieved either by experimentally setting values to λ_t or by using a method that will automatically find this balance as in [7, 36]. We compare fixing λ_t experimentally with different values during training and also use Multi-Term Adam (MTAdam) [34] to find the optimal trade-off between the loss terms. We chose this approach as it is simple to implement and has shown great results. MTAdam is based on the first and second moments of the gradients of the models to balance the optimization of each parameter of the network.

Moreover, we propose another solution to deal with this unbalancing problem, this time by a native adversarial training optic. Since [18], adversarial networks have been known to struggle with instability during training. Some of the most known problems are related to vanishing gradients, as the D converges too quickly and, thus, G generates poor samples as the gradients from D are too small. During our experiments, we noticed this behavior on $\mathcal{M}_{deb}$. As the deblurring task is simpler than semantic segmentation, this task converges rapidly. So, images are not blurred enough to prevent adversarial attacks. To overcome this problem, we propose to adapt the gradient penalty (GP) strategy from [20]. This method enforces the Lipschitz constraint by forcing the norm of the gradients from an adversarial model (D, or $\mathcal{M}_{deb}$) to be 1. Therefore, our objective in Equation 3 becomes:

$$\min_{\theta_{opt}}(\lambda_{seg}\mathcal{L}_{seg}(\hat{y}_i^{seg}, y_i^{seg}) - \lambda_{deb}\mathcal{L}_{deb}(\hat{y}_i^{deb}, x_i) + (\|\frac{\partial\mathcal{L}_{deb}(\hat{y}_i^{deb}, x_i)}{\partial\theta_{opt}}\|_2 - 1)^2). \quad (4)$$

3.3. Implementation details

Datasets. Our experiments make use of Cityscapes [12], a dataset for urban scene understanding. This comprises 5000 densely annotated frames split into 2975 images for training, and 500 for validation, which are used for tests only and not during training. Annotations for the rest of the dataset are not accessible to the public.

Pre-processing. During training, we use only a few data augmentation transformations. We normalize images between 0 and 1 so that each pixel value represents the energy of a point light source. Then we randomly crop the image to a fixed resolution of 512 × 512 pixels and randomly flip it horizontally. We use less data augmentation compared to the original Segformer paper [42] as the optical encoder described in 3.1 applies the PSF based on the geometry of the scene and applies the noise based on the intensity of the image. As such, changing them with further data augmentation would likely result in unrealistic images. During inference, we only normalize the images with the original resolution of 1024 × 2048 to feed the network. For semantic segmentation inference, we follow the original paper instructions by using a sliding strategy to generate the output with the same size as the input.

Metrics. We report semantic segmentation performance using mean intersection over union (mIoU), mean accuracy (mA) ie. the average accuracy of all the classes, and overall accuracy (OA) ie. the accuracy of per pixel classification. We report deblurring quality using peak signal to noise ratio (PSNR) and structural similarity (SSIM).

Implementation details. Optical parameters and deblurring network are trained using Adam [28] with a learning rate of 10^{-3}. The semantic segmentation network is finetuned using AdamW [32] and a learning rate of 6×10^{-5}. We train with batches of 16 images for 200 epochs on a NVIDIA A40 ($\sim$ 1 day). Also, we set $\lambda_{seg} = 1$ for all experiments and we vary $\lambda_{deb} \in \{1, 100, 1000\}$ to understand the effects of increasing the importance of the adversarial examples on the sensor model. These values were chosen experimentally with respect to the difference between the loss values from semantic segmentation and deblurring.

4. Experiments results

In this section we present the results obtained with experimental protocol described in the previous section.

4.1. Utility task

Table 1 displays the semantic segmentation results obtained with various settings. The first row contains the reference to the original implementation from the Segformer paper [42]. This network achieves the best segmentation results as it processes clean images in a higher resolution (1024 × 1024) than in the following experiments, without any blur or noise from the optical encoder. Also, the original training pipeline contains more data augmentation than ours, as explained in Section 3.3. OS denotes the baseline obtained by jointly optimizing the optical encoder and the segmentation network to minimize $\mathcal{L}_{seg}$. It reaches slightly

Table 1. The first seven rows include the comparison between the best results with Segformer-B0 trained with adversarial routine. We use the following notation to identify the training strategies: semantic segmentation (S), optics (O), deblurring (D), and minus (-) means the subsequent task is adversarial. For adversarial training, (MTAdam) indicates that MTAdam was used to balance losses, (GP) indicates that gradient penalty was used, otherwise the fixed value of λ_{deb} is indicated. The last two rows include results from NAFNet used to deblur the privacy images to the LPDR task. ↑ and ↓ indicate values we aim at, respectively maximizing and minimizing for the privacy-preserving objective.

Model	Metrics					
	mA↑	mIOU↑	OA↑	PSNR (dB)↓	SSIM ↓	LPDR Acc. ↓
Original Segformer	-	0.762	-	-	-	43.0%
OS	0.778	0.700	0.935	-	-	-
OS-D (MTAdam)	0.770	0.693	0.935	24.5	0.78	7.7%
OS-D ($\lambda_{deb} = 1$)	0.759	0.680	0.929	16.0	0.758	1.9%
OS-D (GP)	0.758	0.678	0.929	22.7	0.81	0.0%
OS-D ($\lambda_{deb} = 10^2$)	0.744	0.664	0.927	21.1	0.80	0.0%
OS-D ($\lambda_{deb} = 10^3$)	0.713	0.630	0.918	23.2	0.84	0.0%
NAFNet Finetuned	-	-	-	-	-	1.4%
NAFNet Pretrained	-	-	-	-	-	0.0%

worse metrics as it deals with the noise added by the sensor model and the blur from the PSF which impedes segmentation of small objects.

Adversarial training (marked as OS-D) with $\lambda_{deb} = 1$ reduces mIoU by approximately 2% compared to the lens optimized without privacy considerations (OS). This small decrease is explained by the small blur added to the image, which can be seen in Figure 6, as the objective losses are unbalanced. Increasing λ_{deb} to 1000 lowers mIoU by 5% more, as the blur amount increases causing a more significant impact on this task. Still, this has a small effect on the utility task compared to the benefit of improved privacy, as seen by the increased blur amount. Tuning λ_{deb} experimentally to 1000 showed great results, however, it can be time-consuming to try different values to achieve the best results. So, by using MTAdam, we expected the algorithm to find a better trade-off between each gradient's influences on the optical model. However, this technique failed to balance the losses to increase privacy for our dataset. Nevertheless, the gradient penalty (GP) strategy proved to be very efficient and added the desired automation to balance the task's influence on the optical model. By forcing the gradient norm of the deblurring model to 1, it prevents gradient vanishing and, so, the model generates significant adversarial information through training. This result ensures privacy as in OS-D $\lambda_{deb} = 1000$ but shows better performance in semantic segmentation.

Figure 5 shows two examples of results obtained for the utility task with gradient penalty. Compared to the original scene, the sensor image is noticeably blurrier. Fine details such as sign poles are not detected by the semantic segmentation but the quality is good enough to count pedestrians or cars.

Figure 6 displays an example of sensor image for each adversarial system as well as the corresponding PSF and deblurring result. As expected the sensor image becomes blurrier as λ_{deb} increases. The PSF corresponding to the systems with the most blur are darker because their energy is more spread on the sensor.

Overall, the reduction of semantic segmentation metrics seen in Table 1 is compatible with the exploitation of the segmentation results.

4.2. Privacy Attack Experiment

In the following, we study the systems optimized with $\lambda_{deb} = 1000$ and with gradient penalty to assess the potential of privacy attacks. We focus on deblurring attacks that aim at reconstructing sharp images from the sensor, and license plate detection and recognition (LPDR) that would be used to identify vehicles.

Deblurring Attack. We consider two types of deblurring attacks against the co-designed system. First, a sensor access attack, where one only has access to images from the sensor of the privacy-preserving lens. Then, known pairs attack, where one has access to pairs of clean and blurry images. To ensure that the co-designed lens is robust against privacy attacks we test these two scenarios. For the sensor access attack, the lack of ground truth data corresponding to the blurry images means that a deblurring network can't be fine-tuned for the privacy-preserving lens. In this case, we use pretrained weights provided for NAFNet [41] and DeblurGAN-V2 [30] obtained after training on the GoPro dataset. For this attack, we correct the white balance manually to compensate for the tint introduced by the physical sensor model (see Figure 5). For the known pairs attack we train using the parameters recommended for training NAFNet [41] for 100000 iterations over images from the

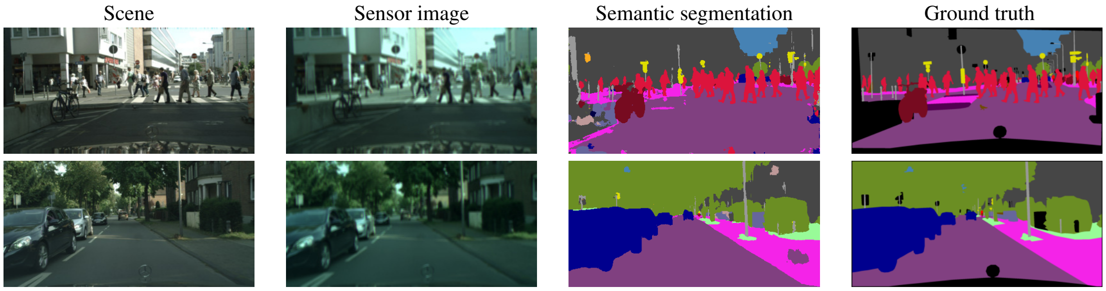

Figure 5. Examples of semantic segmentation results from the adversarial system trained with gradient penalty.

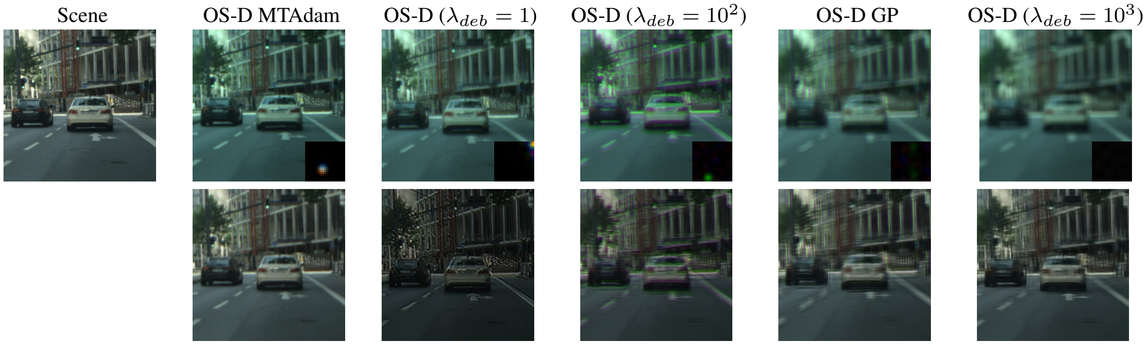
Scene
OS-D MTAdam
OS-D ($\lambda_{deb} = 1$)
OS-D ($\lambda_{deb} = 10^2$)
OS-D GP
OS-D ($\lambda_{deb} = 10^3$)

Figure 6. In the first row, we show examples of sensor images (x_i^s) from the adversarial systems, and their corresponding PSF in the bottom right corner. In the second row, we have the corresponding reconstruction ($\hat{y}_i^{deb}$) from the deblurring network optimized ($\mathcal{M}_{deb}(\cdot\,; \theta_{deb}*)$) during the adversarial training.

Cityscapes dataset after blurring them with the lens optimized for privacy.

Table 2 displays the results of the deblurring attacks: using the deblurring network obtained after the adversarial training; for two networks in the sensor access scenario; and using a network trained on a set of images from the camera with corresponding high quality images. In the first case, both networks trained on the GoPro dataset reach a PSNR and an SSIM value close to the network optimized during the adversarial training. Figure 7 displays examples of images obtained after deblurring for the different attacks. The pretrained networks do not recover high-frequency content in the blurred images making it impossible to see fine details (see Figure 7). On the other hand, the network from the adversarial training introduces some ringing artifacts that lower his PSNR. With access to pairs of blurry and corresponding sharp images, it is possible to train a deblurring network to specifically restore images from our camera. In this case, we reach higher PSNR and SSIM. The resulting images (Figure 7) recover most of the content but some fine details remain indistinguishable.

The resulting camera is robust against deblurring by networks pretrained for generic images but good reconstruction can be achieved when training a dedicated attack network. Note that such an attack would be unlikely in practice since it would require gathering many aligned pairs of blurry and sharp images.

License Plate Detection and Recognition (LPDR). Like in [9], we also consider an LPDR model to simulate an attack on our designed lens. Cityscapes-LP [5] provides an extension to the Cityscapes dataset with license plate annotations for validation.

To simulate this attack, we follow [38] as the proposed method is extensible to license plates from different countries and can be used for inference without further training. The authors propose a pipeline divided into three main parts including a vehicle detection model, LP detection, and finally an OCR (Optical Character Recognition). This last step is a combination of character segmentation and character recognition.

As a metric for this task, we adopt the accuracy (Acc) corresponding to the fraction of correctly detected and recognized license plates over the total number of readable license plates. We express this metric as a percentage. Our results can be observed in the rightmost column of Table 1. We compare the performances obtained with the LPDR model using respectively the original sharp images from Cityscapes dataset (first row) and blurred images from adversarial learning. Moreover, we test the LPDR model on

Deblurring network	OS-D (GP)		OS-D ($\lambda_{deb} = 10^3$)	
	PSNR (dB) ↓	SSIM ↓	PSNR (dB) ↓	SSIM ↓
Adversarial training	22.7	0.81	23.2	0.84
Pretrained NAFNet [41]	25.7	0.80	23.4	0.80
Pretrained DeblurGanV2 [30]	25.8	0.80	23.5	0.84
Trained NAFNet	38.9	0.97	36.7	0.96

Table 2. Comparison of deblurring attacks for the systems trained with gradient policy or with $\lambda_{deb} = 10^3$. The adversarial training corresponds to the RedNet [35] network used during the adversarial training.

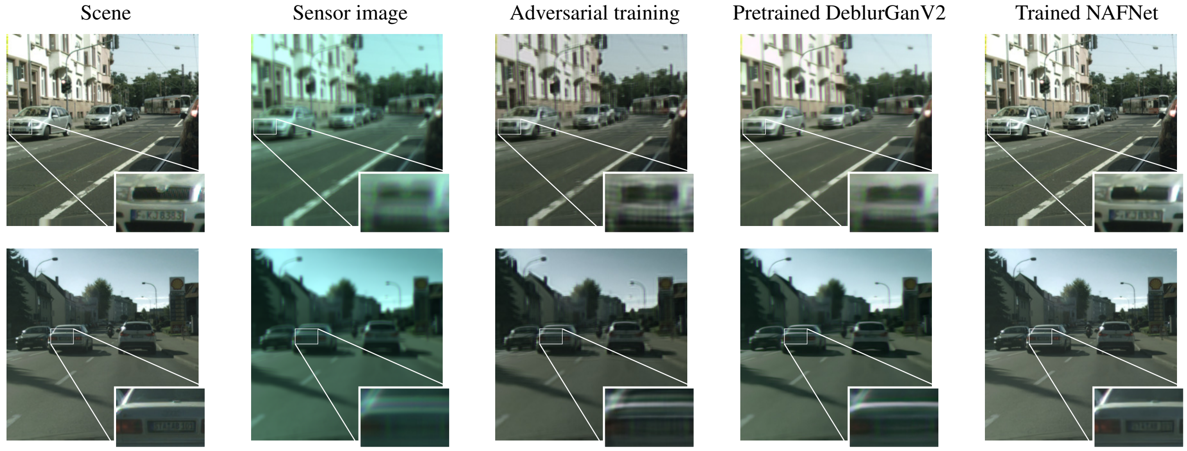

Figure 7. Examples of deconvolved images from the adversarial system trained with gradient penalty. The blue tint of the sensor image comes from different quantum efficiencies for each channel in the sensor model.

the deblurred images obtained using the NAFNet finetuned and pretrained (last two rows). Notice that we aim to have low accuracy to indicate our system is efficient in protecting privacy.

Compared to the performance obtained with the original images, the LRPD model shows drastically reduced scores when applied to the blurred images at the sensor output. This could be explained by the color perturbation added by the sensor that the LPDR model does not expect in addition to the blur. Nevertheless, this score is only slightly improved when the model is applied to the deblurred images using NafNet, which rectifies this effect. To conclude, this experiment validates the gain in privacy protection brought by the proposed method.

5. Conclusions

This paper presents the use of deep-optics tools to improve the privacy of imaging systems by using physical image encoding. We propose an adversarial training procedure to jointly optimize a lens and processing to perform semantic segmentation of urban images while ensuring that access to sensor images would not raise privacy concerns. In this context, we successfully integrate a penalization term for the deblurring model gradients to prevent fast convergence of the adversarial model. Thus, we continually guarantee its feedback to strengthen the optical model against attacks. Besides, our approach ensures privacy in a self-supervised manner and does not require sensitive annotations related to privacy attributes. We evaluate the robustness of the optimized system for two kinds of attacks: deblurring of sensor images and automatic license plate recognition. The proposed tools are generic and could be applied to create privacy-preserving systems for other tasks.

One limitation of our method is the use of Fourier optics which can only simulate PSF on the optical axis. Using differentiable ray tracing will allow us to optimize complex lenses made of multiple glass elements while considering PSF variations in the field. The demonstration of our method has been performed on a simulated scenario. Since, the next step of this project is to design and build a prototype of the privacy-preserving lens and then to validate the method on real-world data.

References

[1] Prachi Agrawal. De-identification for privacy protection in surveillance videos. *Int. Institute of Information Technology*, 2010. 2

[2] Shafiq Ahmad, Pietro Morerio, and Alessio Del Bue. Person re-identification without identification via event anonymization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11132–11141, 2023. 2

[3] Shafiq Ahmad, Gianluca Scarpellini, Pietro Morerio, and Alessio Del Bue. Event-driven re-id: A new benchmark and method towards privacy-preserving person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 459–468, 2022. 2

[4] Naveed Akhtar and Ajmal S. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 4

[5] Saeed Ranjbar Alvar, Korcan Uyanik, and Ivan V Bajić. License plate privacy in collaborative visual analysis of traffic scenes. In *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 300–305. IEEE, 2022. 7

[6] Paula Arguello, Jhon Lopez, Carlos Hinojosa, and Henry Arguello. Optics lens design for privacy-preserving scene captioning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3551–3555, 2022. 2

[7] Julian Blank and Kalyanmoy Deb. Pymoo: Multi-objective optimization in python. *Ieee access*, 8:89497–89509, 2020. 5

[8] Timothy Callemein, Kristof Van Beeck, and Toon Goedemé. How low can you go? privacy-preserving people detection with an omni-directional camera. *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2018. 2

[9] Marcela Carvalho, Oussama Ennaffi, Sylvain Chateau, and Samy Ait Bachir. On the design of privacy-aware cameras: a study on deep neural networks. *Eur. Conf. Comput. Vis. Worksh.*, 2022. 1, 2, 7

[10] J. Chang and G. Wetzstein. Deep optics for monocular depth estimation and 3D object detection. In *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2019. 1

[11] Durkhyun Cho, Jin Han Lee, and Il Hong Suh. Cleanir: Controllable attribute-preserving natural identity remover. *Applied Sciences*, 10(3):1120, 2020. 2

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[13] Geoffroi Côté, Fahim Mannan, Simon Thibault, Jean-François Lalonde, and Felix Heide. The differentiable lens: Compound lens search over glass surfaces and materials for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[14] Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. Spact: Self-supervised privacy preservation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[15] Thomas Dubail, Fidel Alejandro Guerrero Peña, Heitor Rapela Medeiros, Masih Aminbeidokhti, Eric Granger, and Marco Pedersoli. Privacy-preserving person detection using low-resolution infrared cameras. In *European Conference on Computer Vision*, pages 689–702. Springer, 2022. 2

[16] M. Dufraisse, P. Trouvé-Peloux, J.-B. Volatier, and F. Champagnat. Deblur or denoise: the role of an aperture in lens and neural network co-design. *Opt. Lett.*, 48(2):231–234, Jan 2023. 4

[17] Liyue Fan. Image pixelization with differential privacy. In *Data and Applications Security and Privacy XXXII: 32nd Annual IFIP WG 11.3 Conference, DBSec 2018, Bergamo, Italy, July 16–18, 2018, Proceedings 32*, pages 148–162. Springer, 2018. 2

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 4, 5

[19] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company publishers, 2005. 2, 4

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 2, 5

[21] Harel Haim, Shay Elmalem, Raja Giryes, Alex M. Bronstein, and Emanuel Marom. Depth estimation from a single image using deep learned phase coded mask. *IEEE Transactions on Computational Imaging*, 4(3):298–310, 2018. 2

[22] Aymeric Halé, Pauline Trouvé-Peloux, and J-B Volatier. End-to-end sensor and neural network design using differential ray tracing. *Optics express*, 29(21):34748–34761, 2021. 2

[23] Carlos Hinojosa, Miguel Marquez, Henry Arguello, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Privhar: Recognizing human actions from privacy-preserving lens. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 314–332. Springer, 2022. 1, 2, 3

[24] Carlos Hinojosa, Juan Carlos Niebles, and Henry Arguello. Learning privacy-preserving optics for human pose estimation. In *ICCV*, pages 2573–2582, 2021. 2, 3

[25] Håkon Hukkelås and Frank Lindseth. Deepprivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1329–1338, 2023. 2

[26] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *Advances in Visual Computing*, pages 565–578. Springer International Publishing, 2019. 2

[27] Junho Kim, Young Min Kim, Yicheng Wu, Ramzi Zahreddine, Weston A Welge, Gurunandan Krishnan, Sizhuo Ma, and Jian Wang. Privacy-preserving visual localization with event cameras. *arXiv preprint arXiv:2212.03177*, 2022. 2

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[29] Sudhakar Kumawat and Hajime Nagahara. Privacy-preserving action recognition via motion difference quantization. In *ECCV*, pages 518–534. Springer, 2022. 1

[30] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. 6, 8

[31] Zongling Li, Qingyu Hou, Zhipeng Wang, Fanjiao Tan, Jin Liu, and Wei Zhang. End-to-end learned single lens design using fast differentiable ray tracing. *Opt. Lett.*, 46(21):5453–5456, Nov 2021. 2

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 4

[34] Itzik Malkiel and Lior Wolf. Mtadam: Automatic balancing of multiple training loss terms. *arXiv preprint arXiv:2006.14683*, 2020. 5

[35] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 4, 8

[36] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018. 5

[37] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in neural information processing systems*, 32, 2019. 4

[38] Sergio Montazzolli Silva and Claudio Rosito Jung. License plate detection and recognition in unconstrained scenarios. In *Proceedings of the European conference on computer vision (ECCV)*, pages 580–596, 2018. 7

[39] 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 Transactions on Graphics (TOG)*, 37(4):114, 2018. 2

[40] Qilin Sun, Congli Wang, Fu Qiang, Dun Xiong, and Heidrich Wolfgang. End-to-end complex lens design with differentiable ray tracing. *ACM Trans. Graph*, 40(4):1–13, 2021. 2

[41] Xintao Wang, Liangbin Xie, Ke Yu, Kelvin C.K. Chan, Chen Change Loy, and Chao Dong. BasicSR: Open source image and video restoration toolbox. https://github.com/XPixelGroup/BasicSR, 2022. 6, 8

[42] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 4, 5

[43] von F. Zernike. Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode. *Physica*, 1(7):689–704, May 1934. 3