Optimizing Optical and Digital Elements for Privacy Protection in Computer Vision Applications

Sultan Alshehri
Department of Electrical and Computer Engineering
Duke University
Durham, NC
sultan.alshehri@duke.edu

Abstract

With the growing reliance on visual data, the tradeoff between privacy and utility has become a major concern. The need for ensuring personal privacy without sacrificing the value of visual data for machine learning applications has motivated the development of privacy-preserving methods. In this work, we evaluated the general framework of using optical and digital elements to protect privacy in computer vision applications. In particular, a trainable optical filter and a digital encoder were jointly optimized for privacy and utility objectives. We explore two ways of parameterizing the optical filter and compare the results with basic privacy-preserving methods. The method was evaluated on two tasks: image classification and semantic segmentation. The results demonstrate that it can preserve task-relevant information without compromising privacy.

1 Introduction

A growing risk to personal privacy has arisen from the pervasive use of cameras, which exposes people's identities in both public and private contexts such as biometric identification, surveillance, and more. Its use has also become more prevalent in the healthcare domain, with application in patient monitoring [6], surgical skill assessment [15], and systemic disease diagnosis [3]. Meanwhile, the recent progress of deep learning has increased the need for larger datasets, which usually compromise sensitive data and personal information.

A challenging tradeoff between privacy and utility has emerged due to the growing reliance on visual data. On one hand, images provide a rich source for training deep learning models, leading to valuable applications across numerous fields. On the other hand, malicious actors can exploit this data to uncover sensitive information. Several methods have been proposed for privacy protection, including federated learning [4, 14], differential privacy [1, 8], and homomorphic encryption [2]. However, even with advanced data encryption and security methods, the risk of data leakage remains a concern, primarily because data is captured and often stored in the cloud, which is susceptible to attacks. Therefore, methods that are secure and protect privacy have become crucial in computer vision applications. To enable applications that rely on privacy-sensitive data, we investigate the joint optimization of optical and digital elements to protect privacy.

2 Related Work

In this section, we briefly review related work in three areas: privacy protection using traditional methods, privacy-preserving deep learning, and privacy protection using optical and digital systems.

2.1 Privacy Protection using Traditional Methods

Traditional methods have primarily addressed this problem by blurring or obfuscating sensitive information [5, 7, 17]. In [22], the authors assessed two privacy obfuscation methods (blurring and blocking) and their effectiveness in protecting privacy based on human perception. These methods are mainly used to distort images using hand-crafted techniques such as blurring or downsampling, but often result in removing useful features and adversely affecting the performance of downstream tasks.

2.2 Privacy Protection using Deep Learning

Many recent proposed methods have investigated privacy in deep learning. The work by [12] proposed an optimization framework that trains a privacy network and an adversarial classifier to privatize the data. In [10], the authors train an adversarial obfuscator network to remove task-irrelevant image features to preserve privacy. Similarly, [19] proposed a method that eliminates specific attributes to prevent identification while preserving task-specific information using an adversarial training approach. Other methods, such as [13], use inpainting techniques to de-identify faces while keeping the original data distribution.

2.3 Deep Computational Optics for Privacy Protection

The advancement in deep computational optics has motivated the integration of optical elements into the deep learning pipeline to protect image privacy. By incorporating an optimizable optical encoder, several methods, such as [11, 20], have been proposed to learn a privacy-preserving optical filter for tasks such as pose and depth estimation while degrading personal human attributes.

3 Methods

3.1 Trainable Optical Filter

The training objective is to enhance privacy by preventing sensitive information from being inferred or recovered from images while preserving their utility for downstream tasks.

3.1.1 Direct PSF

Let the input image be denoted as $I \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and channels of the image, respectively. To simulate the physical process of light passing through an optical system, the input image is convolved with a point spread function (PSF) kernel. In our Direct PSF approach, we parameterize the PSF as a trainable convolutional layer $H \in \mathbb{R}^{h \times w \times C}$. The details of these kernel parameters will be further discussed in the Experimental Evaluation section.

We impose certain constraints on the PSF kernel to account for the physical limitations of the optical element. Firstly, we optimize the kernel weights with a non-negative constraint. Subsequently, the weights are normalized to satisfy the following condition:

$$\sum_{i=1}^{h} \sum_{j=1}^{w} H_c(i,j) = 1 \tag{1}$$

Where c denotes the RGB channels of the image.

The kernel is then convolved as a depthwise convolution with the input image as follows:

$$I_c^{'} = I_c * H_c, \quad \Rightarrow I_c^{'}(x,y) = \sum_{i=1}^{h} \sum_{j=1}^{w} I_c(i,j) \cdot H_c(x+i, x+j), \quad \forall c \in C \tag{2}$$

Where h and w are the height and width of the kernel, and $I_c^{'}$ is the output of the optical kernel for the c^{th} channel. The output is then fed into a convolutional neural network (CNN), which acts as a digital encoder that can effectively learn from the distorted representations of the PSF-convolved images to perform downstream tasks.

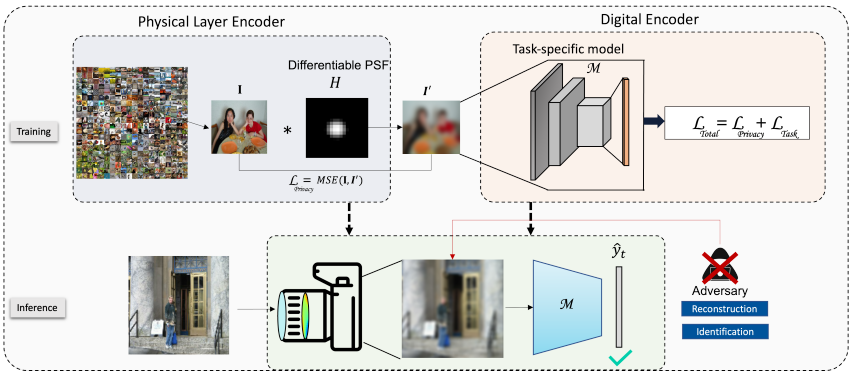


Figure 1: Overall project workflow.

3.1.2 PSF using Zernike Polynomials

The refractive optical encoder can be parameterized using Zernike polynomials, where the coefficients of the Zernike basis can be learned to optimize the optical kernel of the lens. Zernike polynomials describe the wavefront aberration of the lens using a set of orthogonal polynomials [18] as follows:

$$\phi(x, y) = \sum_{i=1}^{N} a_i Z_i(x, y) \tag{3}$$

Here, Z_i is the i^{th} Zernike polynomial with Noll index i, and a_i is the corresponding coefficient. By expressing the optical kernel as a linear combination of Zernike polynomials, we can learn a set of coefficients optimized for our objective function, resulting in a smoother kernel for the optical encoder [11].

3.2 Loss Function

In training the optical kernel, our goal is to learn privacy-utility loss that can be used for downstream tasks. We define the loss as a combination of the privacy loss and the task loss:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda \mathcal{L}_{privacy} \tag{4}$$

Where λ is a hyperparameter that is chosen based on the application. In the task loss for image classification, we use the categorical cross-entropy (CE) loss:

$$\mathcal{L}_{CE} = -\sum_{i} y_i \log(\hat{y}_i) \tag{5}$$

Where $\hat{y}_i$ is the output of a softmax layer, and y_i is the ground truth label. For the segmentation task, the Dice loss is used as the task loss:

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_i y_i \hat{y}_i}{\sum_i y_i + \sum_i \hat{y}_i} \tag{6}$$

For the privacy loss, the objective is to minimize the sensitive information content in the output of the optical layer, and it is achieved by learning a mapping that distorts the image while preserving the information relevant for downstream tasks. We define the privacy loss as maximizing the Mean Squared Error (MSE) between the input image and the output of the PSF layer. From our experiments, we found that the MSE loss provides a good balance for privacy. A regularization term can be added to the loss function by computing the Total Variation (TV) over $I^{'}$, which has been shown to enhance the privacy by reducing high-frequency components in the output image. The TV term penalizes sharp transitions in the image, thus promoting a more blurred output. The privacy loss is defined as

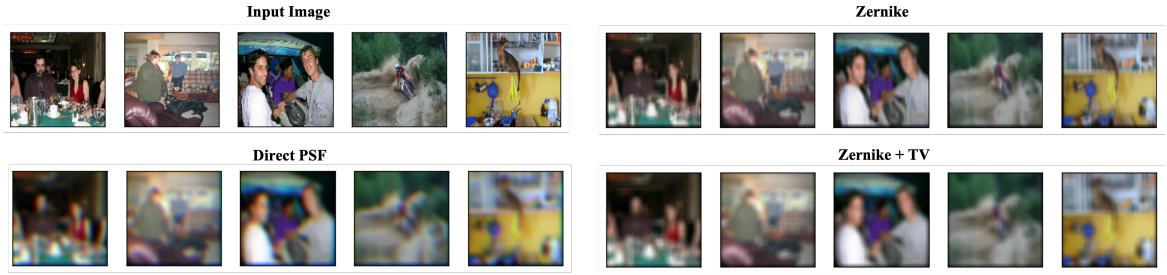


Figure 2: Qualitative comparison between the input image (standard lens) and the output of the optimized lens.

follows:

$$\mathcal{L}_{privacy} = \lambda \mathcal{L}_{MSE} + \beta \mathcal{L}_{TV} \tag{7}$$

$$= -MSE(I, I^{'}) + TV(I^{'}) \tag{8}$$

$$= -\frac{1}{HWC} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} (I_{h,w,c} - I^{\prime}_{h,w,c})^2 + \sum_{c} \left(\sum_{h,w} |\Delta_h I^{'}_{h,w,c}| + \sum_{h,w} |\Delta_w I^{'}_{h,w,c}| \right) \tag{9}$$

Where I is the original image, and $I^{'}$ is the output of the physical layer.

4 Experimental Evaluation

4.1 Datasets

We trained our models on the Pascal VOC 2012 [9] and used the 20 classes for the classification task. The Common Objects in Context (COCO) [16] was employed for evaluating the model inversion attack. The segmentation task was evaluated on [21], which consists of 5678 images and its semantic binary masks. We have consistently used 0.9/0.1 train/test splits for all the datasets. In addition, the images were resized to 128x128 in all the experiments to fit the computational resources.

Method	SSIM ↓	MS-SSIM ↓	PSNR ↓	Acc (Top-1 %) ↑	Acc (Top-3 %) ↑
Without PSF	-	-	-	93.07	99.451
Blur Kernel($\sigma = 2$)	0.5742	0.8923	21.5892	89.95	94.63
Blur Kernel($\sigma = 5$)	0.3927	0.7062	18.8584	74.78	91.07
Gaussian Noise ($\sigma = 0.1$)	0.2913	0.7563	15.0149	82.06	92.89
Downsample (x6)	0.4756	0.7814	19.3893	89.60	94.97
Direct PSF	0.4235	0.7414	17.5525	92.80	95.41
Zernike PSF	0.3446	0.6515	16.0731	93.41	95.84
Zernike PSF (TV)	0.4248	0.7726	18.3925	91.24	95.67

Table 1: Evaluation results of the classification task, benchmarked against non-private standard lens and different naive privacy-preserving methods.

4.2 Network and Training

We used Tensorflow for conducting our experiments. In the classification task, we trained a Convolutional Neural Network (CNN) consisting of three convolutional layers and two dense layers with a

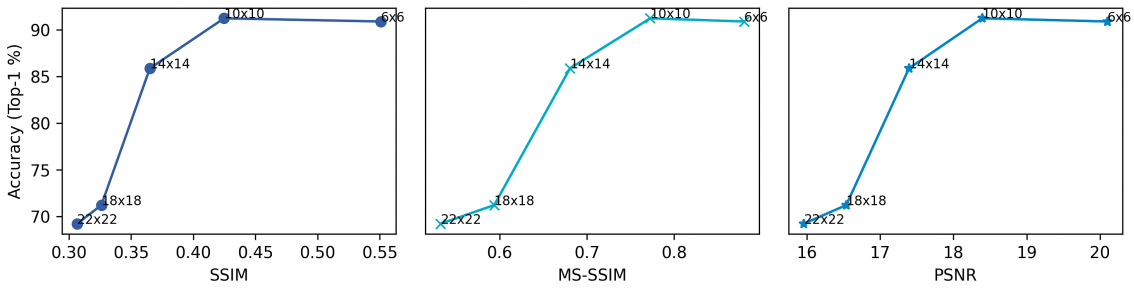


Figure 3: Tradeoff between privacy and utility on three evaluation metrics using four kernel sizes.

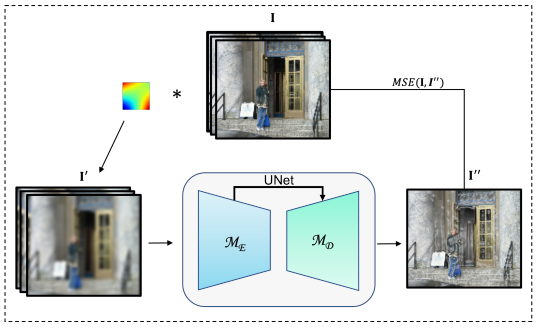
softmax activation function. For the direct PSF approach, we parameterized the PSF as a trainable convolutional kernel, which was incorporated as the first layer in the CNN. We assigned a kernel size of 22x22 and initialized the weights from a Gaussian distribution. For the Zernike PSF approach, we used the first 6 Zernike polynomials to parameterize the optical kernel and have set the Zernike modes radius to 5.

Consistent hyperparameters were used across all experiments to ensure a fair comparison between the various methods. The model was trained for 30 epochs using the Adam optimizer, a batch size of 32, and a learning rate of 0.001. The learning rate was updated using an exponential decay schedule, with a decay factor of 0.1 applied after 20 and 25 epochs.

For the segmentation task, we adopted a U-Net architecture with a MobileNetV2 encoder pretrained on ImageNet. The decoder consisted of four upsampling blocks, with each block including a transposed convolution layer, a batch normalization layer, and a ReLU activation. The images and masks were augmented using random horizontal flips, channel shuffling, and rotation, resulting in a total of about 30k images. The model was trained for two epochs using the Adam optimizer, with a batch size of 2 and a learning rate of 0.001.

4.3 Privacy and Task Metrics

To evaluate the performance of the used methods, we employed three perceptual metrics as a measure of privacy, in addition to the classification accuracy. The Structural Similarity Index (SSIM), its variant, the Multi-Scale Structural Similarity Index (MS-SSIM), and the Peak Signal-to-Noise Ratio (PSNR) were used to provide a quantitative measure of the similarity between the original image and the image after applying the PSF. The classification task was evaluated based on the Top-1 and Top-3 accuracy, while the Intersection over Union (IoU) was used to evaluate the segmentation task.



(a) Workflow of reconstructing the image from the PSF-convolved image.

(b) Samples of the reconstructed images.

Figure 4: Privacy inversion attack.

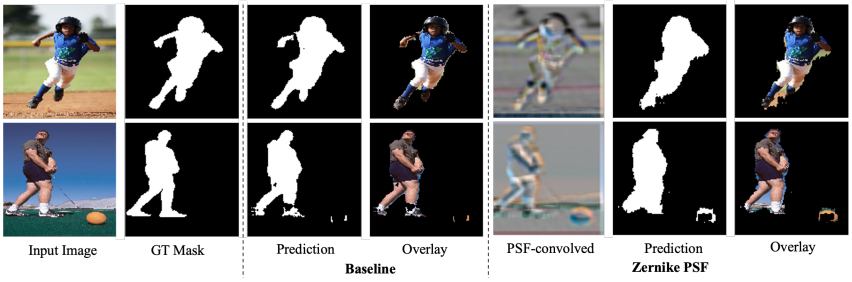


Figure 5: Qualitative comparison of the predicted masks using the baseline and the PSF approach.

Model (Configuration)	SSIM ↓	MS-SSIM ↓	PSNR ↓	IoU ↑
UNet	-	-	-	0.7787
UNet (Zernike PSF)	0.2303	0.1105	10.0892	0.6296

Table 2: Evaluation results of the segmentation task.

5 Results and Discussion

5.1 Image Classification

Figure 2 shows few samples of the convolved images using different PSFs. Compared to the Zernike PSF, the TV regularized Zernike PSF resulted in a more blurred image. Table 1 shows the results of the image classification task, where it is benchmarked against different handcrafted augmentation methods simulating simple privacy-preserving strategies. While the basic methods such as blurring and downsampling can produce images with high privacy according to the SSIM, MS-SSIM, and PSNR measures, their performance in the downstream task is noticeably affected. It can be shown that the optimized PSF methods produce a comparable or better privacy level while still maintaining a close classification accuracy to the baseline approach. This indicates that the PSF approach retained the task-relevant information essential for performing the downstream task while optimizing for the privacy loss.

Figure 3 demonstrates the privacy-utility trade-off for four different kernel sizes on the perceptual metrics. We can observe that as the kernel size increases, the degree of image distortion becomes more significant, which leads to better privacy protection. However, the classification performance also decreases as a result of the increasing distortion. Therefore, the choice of kernel size is important in determining the degree of sensitive information in the image without compromising the ability of the model to effectively perform the task.

Another experiment was performed to assess the robustness of the method against the privacy inversion attack. In this experiment, assuming an adversary has black-box access to the optimized PSF, we trained a UNet model to deblur the convolved images. In Figure 4a, we used an optimized PSF to simulate a new dataset of convolved images using the COCO dataset. We then trained a UNet model to deblur the images to recover the original images. Visual inspection of the results in Figure 4b shows that the attempt to recover the original images was unsuccessful.

5.2 Semantic Segmentation

Comparable results were obtained for the semantic segmentation task. In table 2, when comparing the results of the trained U-Net model with the designed PSF to the baseline U-Net model, we find that the SSIM, MS-SSIM, and PSNR are remarkably lower. However, the reported Intersection over Union (IoU) metric shows a slight decrease with the PSF approach. Considering that the model was trained for two epochs, we expect better performance with more training and hyperparameter tuning. Figure 5 offers a qualitative comparison between the baseline and the Zernike PSF approach. The PSF-convolved images show a significant distortion while preserving the output semantic edges, which is essential for performing segmentation tasks.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345, 2017.

[3] Boris Babenko, Ilana Traynis, Christina Chen, Preeti Singh, Akib Uddin, Jorge Cuadros, Lauren P Daskivich, April Y Maa, Ramasamy Kim, Eugene Yu-Chuan Kang, et al. A deep learning model for novel systemic biomarkers in photographs of the external eye: a retrospective study. *The Lancet Digital Health*, 2023.

[4] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1:374–388, 2019.

[5] Michael Boyle, Christopher Edwards, and Saul Greenberg. The effects of filtered video on awareness and privacy. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 1–10, 2000.

[6] Anis Davoudi, Kumar Rohit Malhotra, Benjamin Shickel, Scott Siegel, Seth Williams, Matthew Ruppert, Emel Bihorac, Tezcan Ozrazgat-Baslanti, Patrick J Tighe, Azra Bihorac, et al. Intelligent icu for autonomous patient monitoring using pervasive sensing and deep learning. *Scientific reports*, 9(1):1–13, 2019.

[7] Benedikt Driessen and Markus Dürmuth. Achieving anonymity against major face recognition algorithms. In *Communications and Multimedia Security: 14th IFIP TC 6/TC 11 International Conference, CMS 2013, Magdeburg, Germany, September 25-26, 2013. Proceedings 14*, pages 18–33. Springer, 2013.

[8] Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings 5*, pages 1–19. Springer, 2008.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[10] Aayush Gupta, Ayush Jaiswal, Yue Wu, Vivek Yadav, and Pradeep Natarajan. Adversarial mask generation for preserving visual privacy. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE, 2021.

[11] Carlos Hinojosa, Juan Carlos Niebles, and Henry Arguello. Learning privacy-preserving optics for human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2573–2582, 2021.

[12] Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. Generative adversarial privacy. *arXiv preprint arXiv:1807.05306*, 2018.

[13] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, 2019, Proceedings, Part I 14*, pages 565–578. Springer, 2019.

[14] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[15] Kyle Lam, Junhong Chen, Zeyu Wang, Fahad M Iqbal, Ara Darzi, Benny Lo, Sanjay Purkayastha, and James M Kinross. Machine learning for technical skill assessment in surgery: a systematic review. *NPJ Digital Medicine*, 5(1):24, 2022.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer*

Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014.

[17] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

[18] Robert J Noll. Zernike polynomials and atmospheric turbulence. *JOsA*, 66(3):207–211, 1976.

[19] Francesco Pittaluga, Sanjeev Koppal, and Ayan Chakrabarti. Learning privacy preserving encodings through adversarial training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 791–799. IEEE, 2019.

[20] Zaid Tasneem, Giovanni Milione, Yi-Hsuan Tsai, Xiang Yu, Ashok Veeraraghavan, Manmohan Chandraker, and Francesco Pittaluga. Learning phase mask for privacy-preserving passive depth estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 504–521. Springer, 2022.

[21] Nikhil Tomar. Person segmentation. `https://www.kaggle.com/datasets/nikhilroxtomar/person-segmentation`, 2021. Retrieved on 10 April 2023.

[22] Nishant Vishwamitra, Bart Knijnenburg, Hongxin Hu, Yifang P Kelly Caine, et al. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 39–47, 2017.