Supplementary Material: Privacy-preserving Optics for Enhancing Protection in Face De-identification

Jhon Lopez[1,2], Carlos Hinojosa[2,*], Henry Arguello[1,†], Bernard Ghanem[2,†]
[1]Universidad Industrial de Santander; [2]King Abdullah University of Science and Technology (KAUST)
https://carloshinojosa.me/project/privacy-face-deid/

Introduction

This supplementary document provides additional experiments, analysis, and discussion of our work. Specifically, this document includes the following:

1. Light propagation and image formation model.
2. Proposed regularizer in our optics loss function.
3. Training details.
4. Experiments using latent vectors to generate the style codes.
5. Experiments using the source image in the discriminator.
6. Additional ablation study: End-to-end training.
7. PSF frequency analysis.
8. Non-blind deconvolution attack.
9. Human evaluation study details.
10. Qualitative results using low-resolution cameras.
11. Additional qualitative results using our camera.
12. Additional hardware experiments.
13. Concluding remarks.
14. Personal data/human subjects discussion.

1. Light propagation and image formation model

We adopt the image formation model previously established in other works [2, 7, 15]. More precisely, our approach involves modeling the light transport within the camera using a differentiable Fourier optics framework [6]. Our optical system comprises a camera with two thin convex lenses and a phase mask (ϕ) between them; see Fig. 1. Considering a thin lens with a focal length f located at a distance d_2 from the sensor, the relationship between the in-focus distance and the sensor distance in the paraxial ray approximation is given by the thin-lens equation: $1/f = 1/d_1 + 1/d_2$. Therefore, an object at a distance d_1 in front of the lens appears in focus at a distance d_2 behind the lens. We first propagate the light emitted by the point, represented as a spherical wave, to the lens. The complex-valued wave field immediately before the lens is given by:

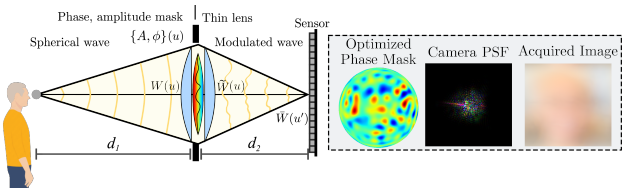

Figure 1. Schematic diagram of the light propagation from an object at a distance d_1 of the lens to the sensor with the focal length d_2. The phase of the spherical light wave coming from a scene point is modulated by our optimized phase mask and captured by the camera's sensor. We take the magnitude-square of the light intensity measured by the sensor to find the values of the PSF. As a result, Our camera captures privacy-preserving images.

$$W(u,v) = \exp\left(ik\frac{u^2+v^2}{d_1}\right)$$

where $k = 2\pi/\lambda$ is the wavenumber. The refractive optical element first delays the phase of this incident wavefront by an amount proportional to the phase mask ϕ of the optical element at each point (u, v). Equivalently, this phase transformation can be mathematically represented as

$$t_{\phi}(u,v) = \exp(ik(n(\lambda)-1)\phi(u,v)),$$

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

The light wave continues to propagate to the camera lens, which induces the following phase transformation [6]

$$t_L(u,v) = \exp\left(-i\frac{k}{2d_1}(u^2+v^2)\right).$$

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

$$\tilde{W}(u,v) = A(u,v)t_{\phi}(u,v)t_L(u,v)W(u,v).$$

*Project lead; † Equal PI contribution.

Finally, the field propagates a distance d_2 to the sensor with the transfer function [6]:

$$T(f_u, f_v) = \exp\left[ikd_2\sqrt{1-(\lambda f_u)^2-(\lambda f_v)^2}\right],$$

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

$$\bar{W}(u', v') = \mathcal{F}^{-1}\left\{\mathcal{F}\left\{\tilde{W}(u, v)\right\} \cdot T(f_u, f_v)\right\},$$

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

$$H(u', v') = |\bar{W}(u', v')|^2.$$

2. Proposed regularizer in our optics loss function

In this work, we propose regularizing the PSF during training to promote symmetric and low frequencies using a predefined defocus regularizer. Specifically, our regularizing approach ensures the obtained PSF is feasible for fabrication and implementation in a laboratory setting and provides faster and more effective convergence during optimization. Our approach represents an improvement over previous methods, which suffer from unstable training and convergence difficulties [7]. We perform an ablation study of our proposed regularizer $\mathbf{H}_f$ to show its relevance in our proposed framework. The PSFs and the privacy-preserving images obtained from this experiment are shown in Fig 2.

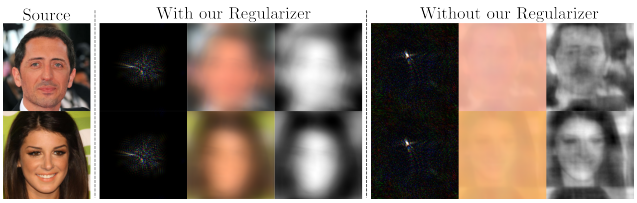

Figure 2. Ablation study of our proposed regularizer $\mathbf{H}_f$. Using our proposed regularizer leads to faster convergence and more stable training.

As observed in Fig 2, the absence of our regularizer results in a PSF that is larger and off-center along the sensor, leading to a noticeable displacement in the captured image; see the grayscale image corresponding to the blue channel. Also, as observed from this grayscale image, the distortion level achieved when not using our regularizer is significantly less than when using it. Therefore, if our regularizer is not used, our framework requires additional hyperparameter tunning and training for more epochs.

3. Training details

Our proposed approach was implemented using PyTorch, and the models were trained in a single NVIDIA A100 GPU. In the first stage, for training the heatmap regression network and the Zernike camera parameters, we use Adam optimizer [11] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rates are set to 5×10^{-4} for the two models. We perform the training for 10 epochs using the FFHQ dataset [9], with a batch size of 32. We use a pre-trained heatmap regression model $\mathcal{U}^*$ to obtain the ground truth for each face [16]. On the other hand, it is important to note that we do not train our regression model $\mathcal{U}$ from scratch but finetune from the pre-trained weights. Additionally, we assume an aberration-free lens as a starting point for start training our optics, i.e., the camera parameters are configured such that the acquired image has no distortion at the beginning of the training. During our training, we finetune the camera model and the regression network $\mathcal{U}$ parameters end-to-end to gradually distort the optics but preserve useful information to extract the heatmap.

In the second stage, to train the generative network, we set the batch size to 8 and the loss functions hyperparameters with the next values: $\lambda_{sty} = 10, \lambda_{ds} = 4, \lambda_{cyc} = 5, \lambda_{LPIPS} = 2 \times 10^2, \lambda_{expr} = 2$. During the training, we use the pre-trained weights provided by the StarGAN2 authors [3] and use their same approach to decrease the loss weight λ_{ds} to zero over the 100k iterations linearly and adopt the non-saturating adversarial loss [5] with $R1$ regularization [13]. We use Adam optimizer [11] with $\beta_1 = 0$ and $\beta_2 = 0.99$. The learning rates, for $\mathcal{G}, \mathcal{S}, \mathcal{E}$, and $\mathcal{D}$ are set to 10^{-4}. During the evaluation, we follow the same approach presented in [3, 8, 17] and use exponential moving averages for all modules but the discriminator $\mathcal{D}$.

4. Experiments using latent vectors to generate the style codes

To inject a specific style into our de-identified images, we adopt StarGAN2 as our generative model. StarGAN2 can extract and generate styles from a reference image and latent vectors. In the main paper, we present our results when the style is extracted from a reference image. In this supplementary document, we include some experiments using the mapping network module presented in StarGAN2 to generate a style vector and evaluate how our method performs using such style vectors. Specifically, we load the pre-trained mapping network and train this module with a learning rate of 10^{-6}, jointly with our generator during 100k iterations; some results of this experiment are shown in Fig. 3. As observed in this figure, the quality of the generated faces is low compared to our results presented in the main document. Also, we observed that having control over the style generation in the de-identified image is challenging. Although the results are not better than those using reference images, from this experiment, we show that our proposed framework can work with the style extracted from latent vectors; hence, additional reference images may not be necessary.

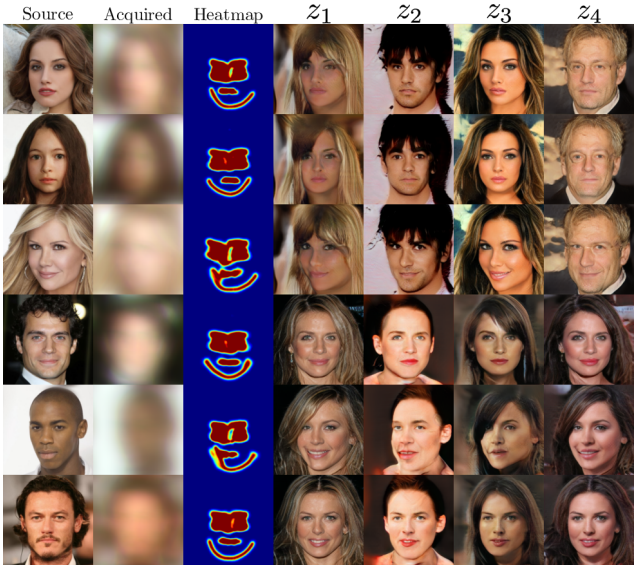

Figure 3. Latent-guided synthesis generation on CelebA-HQ.

5. Experiments using the source image in the discriminator

In contrast to StarGAN2, which trains its discriminator using source images, our discriminator is trained with reference images as we aim to prevent filtering facial information from the source image during the optimization process. To assess the effectiveness of this approach, we conducted ablation studies similar to those presented in the main paper. Specifically, we analyzed the impact of our cost functions to enhance task performance while employing the original image in the discriminator. Table 1 presents the quantitative results of our ablation studies, where the discriminator is trained using the source image. As the table shows, image generation quality improves significantly compared to the results in Tab. 1 of the main document. When the reference image is used to train the generative model, our proposed approach achieves landmark detection metrics close to 2. However, when the original source image is used, the values are close to 1.65, representing an accuracy increase of up to 17%. This trend is also observed in the Bounding Box and FID metrics. On the other hand, the metrics that measure the quality of face de-identification are lower by up to approximately 29% when the source image is used in the discriminator. These results show that facial details are filtered through the discriminator to the generator during the generative model training, possibly due to the $R1$ regularizer.

6. Additional ablation study: end-to-end training

In our main document, we introduce our proposed framework, which employs a two-stage training scheme. This approach was chosen for its more stable optimization and faster convergence in terms of privacy preservation and face de-identification metrics. In this supplementary document, we perform an ablation study on training our proposed framework using an end-to-end (E2E) optimization approach during 100k iterations. Fig. 4 shows examples of the source, acquired, output, and reference images from a batch of the first and the final iterations when training the E2E model. The figure shows that the module convergence is slow, the distortion level is insufficient, the privacy is not well preserved in the image, and the facial heatmap is inaccurate. As we show in the main document, the information provided by the heatmap is important to guide the correct generation of the face position within the de-identified image. These results highlight the importance of two-stage training in our proposed framework.

Components			DIS↑			Landmarks ↓		Bounding Box ↓		FID↓
H	L	E	FR	CASIA	VGGFace2	MtCNN	Dlib	MtCNN	Dlib	
✓	✓	×	0.544	**0.739**	0.848	1.538	1.324	3.894	3.025	**18.548**
✓	×	✓	0.526	0.693	0.793	**1.467**	**1.268**	**3.759**	**2.995**	19.713
✓	✓	✓	**0.598**	0.733	**0.880**	1.691	1.644	4.431	3.751	22.276

Table 1. Quantitative results of the ablation studies using the source image in the discriminator. The best result for each pretrained face recognition model is in bold, and the second-best result is underlined.

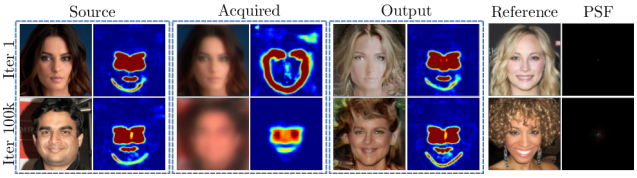

Figure 4. Samples from a batch of the first and the last iteration during the end-to-end training ablation study.

7. PSF frequency analysis

We additionally validate our proposed PSF using the modulation transfer function (MTF) metric [1]. The MTF is computed as the radially averaged magnitude spectrum of the PSF. As observed in Fig. 5, the magnitude spectrum of the proposed PSF decreases significantly for the entire frequency range, indicating low invertibility characteristics, especially in the high-frequency range. This explains why our privacy-preserving images are more robust against deconvolution attacks than low-resolution or defocus cameras. Also, observe that the simulated PSF has lower invertibility than calibrated PSF. This is expected due to our real system's limitations.

8. Non-blind deconvolution attack

In this section, we investigate the robustness of our acquired privacy-preserving images acquired with our optimized camera to deconvolution attacks. In general, there are

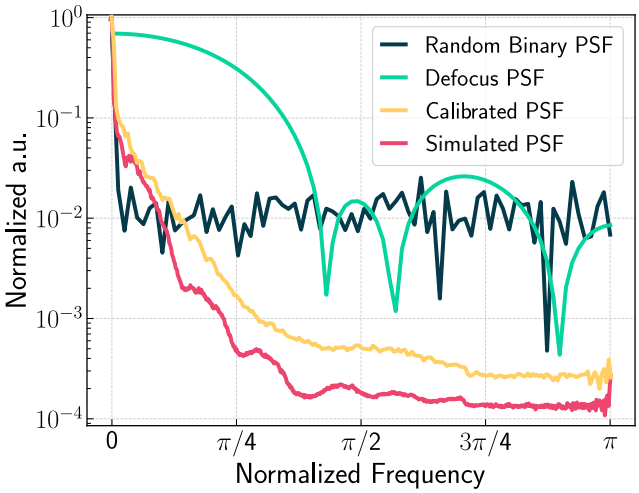

Figure 5. Modulation Transfer Function (MTF) [1] of the point-spread functions (PSFs). The MTF is computed as the radially averaged magnitude spectrum of the PSF. The PSFs compared are Defocus PSF [14], random binary, our simulated PSF, and our calibrated PSF of our proof-of-concept system. The magnitude spectrum of the proposed PSF decreases significantly for the entire frequency range, indicating low invertibility characteristics, especially in the high-frequency range.

two scenarios: in the worst scenario, an attacker has access to the camera and knows the set of Zernike coefficients that form the surface profile ϕ, i.e., the PSF is known. Then, the attacker could perform a **non-blind** deconvolution to reveal the identity of a person within the scene. In the other scenario, an attacker can access a large collection of blur images acquired with our proposed camera but does not know the PSF and can train a **blind** deconvolution network. We explore both scenarios (blind and non-blind deconvolution) and show the results in Fig. 6 and Fig. 7.

To test the robustness of our designed lens to blind deconvolution attacks, we trained a deconvolution network (DeblurGAN [12]) with 28000 sharp and blur (ours) images/frames from the CelebA-HQ dataset acquired with our camera. We use the same default parameters for DeblurGAN and train the network during 300 epochs. For testing, we use the images from FFHQ dataset. Reconstruction is challenging as observed from the results in Fig. 6. The network can reconstruct some objects; however, the face details are missed, and the network cannot recover people's identities.

On the other hand, we also use a non-blind deconvolution approach (Wiener deconvolution [4]) to try to recover the underlying scene. This is the worst scenario for our method since we are assuming an attacker has direct access to the camera. As observed from results in Fig. 7, this approach works better than DeblurGAN as it assumes the PSF is known beforehand. Although this deconvolution approach could retrieve some information from the source

Figure 6. Blind deconvolution of privacy-preserving images acquired by our camera using DeblurGAN-v2 [12].

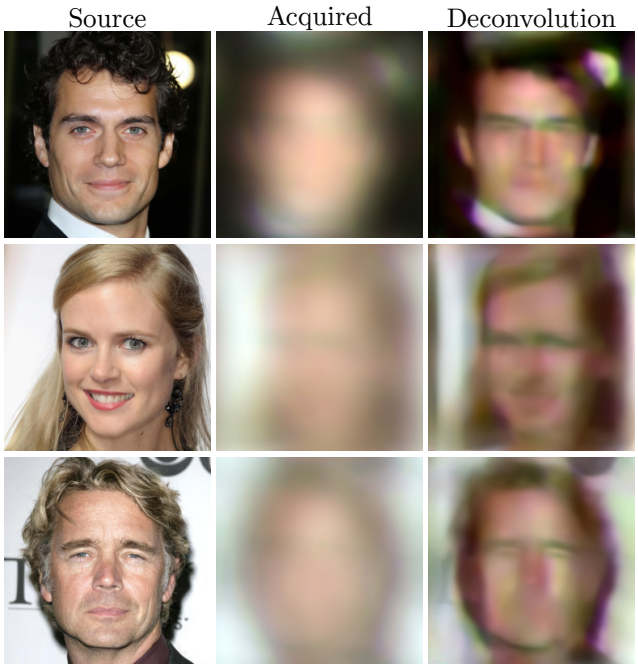

Figure 7. Non-blind deconvolution of privacy-preserving images acquired by our camera using Deep Wiener Deconvolution [4].

face, it is still difficult to recognize the underlying person from the deconvolution result.

Note that the deconvolution attack described in the main document corresponds to a more realistic scenario where the attacker performs a sniffing attack and gets access to one privacy-preserving image being transferred on the internet. In that scenario, the attacker can only leverage a state-of-the-art deconvolution attack to try to recover the identity of the person. For this scenario, we use the DDRM algorithm [10], which leverages the power of several pre-trained diffusion models to solve different inversion problems like deblurring, super-resolution, etc. Note that there is no training script for DDRM since they propose a plug-and-play method that only works with pre-trained diffusion models.

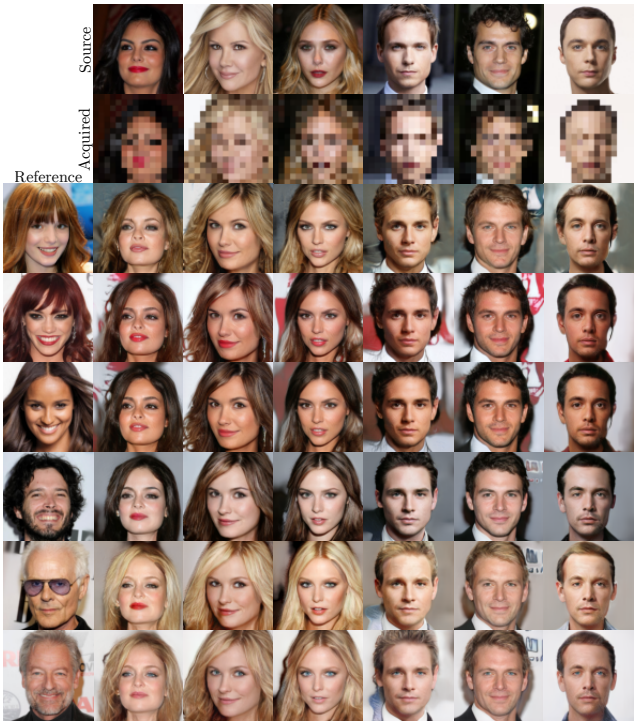

Figure 8. Qualitative results of the proposed method using a low-resolution privacy version on CelebA-HQ dataset.

9. Human evaluation study details

In our research, we conducted a human evaluation to assess the effectiveness of our privacy protection approach under different scenarios. Figure 9 presents the results from a survey conducted among 106 individuals from different ranges of ages, academic backgrounds, and professional fields. Our survey is divided into five main sections. The results from the first section were presented in the main document. In the figure, we show the results for sections 2-5. Each plot presents the answers to one section of questions. Figure 9 (a) presents the results of three questions. For each question, we generate a new face image with our proposed face de-identification framework and show it alongside the other five face images, where one of them corresponds to the current person we want to protect (source image). Then, we asked the individuals to identify the source image.

Section (b) (Figure 9 (b)) is similar to (a), but here we show the privacy-preserving image acquired by our camera alongside the other five non-blurred face images. Then, we asked the individuals to identify the source image. Section (c) (Figure 9 (c)) is similar to (b), but here we show a low-resolution image instead. Finally, in section (d) (Figure 9 (d)), we show four pairs of images and ask participants to determine if one of the images is fake, both images are real, or both are fake. This section aims to assess the quality of the images generated by our proposed method.

In most cases, subjects responded incorrectly, indicating that the privacy-preserving images acquired by our camera effectively preserve privacy by preventing accurate subject identification. Additionally, the de-identified faces generated by our proposed method lead individuals to fail to identify the original subject, which shows our method works as expected. It is also important to note that success rates are higher in Section (c), where participants attempt to identify the subject from a low-resolution image. In the following link, we publish the conducted survey, where readers can explore the questions from the different sections: https://forms.gle/DQZTA7QibaJ45gGH9. To further support our research, we invite the readers to participate in our survey.

10. Qualitative results using low-resolution cameras

In the main document, we present quantitative results for an additional privacy-preserving approach named low-resolution (Ours-LR). Figure 8 shows qualitative results obtained from this approach. We show the source image and their respective low-resolution version in the first two rows. Our generative model generates new faces using the low-resolution image version and various reference images shown in the first column. While the face de-identification performance is accurate, it finds challenges in effectively transferring the reference style to the de-identified image. In some cases, the generated faces retain the style of the source image or produce a new style that differs from both the source and reference images. This is because the low-resolution image preserves high information about the style from the source image.

11. Additional qualitative results using our camera

Fig. 10 shows the visual results of the proposed method applied to face video frames using different references. These qualitative results show the capability of our method to conceal the true identity of the source but maintain global geometry (e.g., head position) in the generated faces. However, the generative model may fail when the face has significant rotation. Additionally, Figure 11 shows some results of rotated faces from the CelebA-HQ dataset where our generative model outputs can fail or introduce visual artifacts. This limitation will be explored in future works.

12. Additional hardware experiments

To complement the results obtained from our real camera prototype, we conducted an experiment using ground truth images as a reference. Specifically, we employed our best model as a starting point and fine-tuned it for 30k iterations with a learning rate of 5×10^{-5}. Figure 12 shows

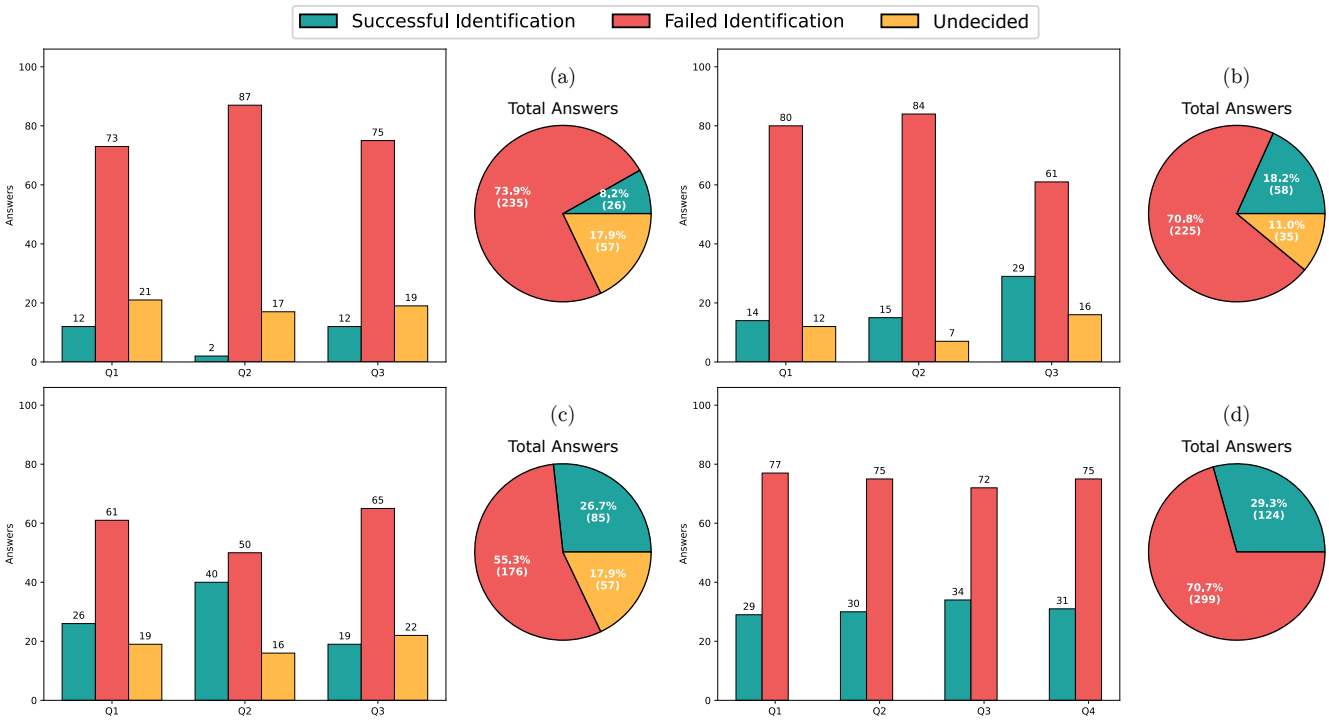

Figure 9. Human evaluation results. (a), (b), (c), and (d), show the results from sections 2, 3, 4, and 5 of our study, respectively. The results from section 1 were presented in the main document. In the figure, Q1, Q2, Q3, and Q4 denote the questions in each section.

Figure 10. Visualization of face de-identification performance on an internet video.

Figure 11. Examples of failure cases generated by our proposed approach on the CelebA-HQ test set.

the results of this experiment. As seen in this figure, when ground truth faces are used as a reference, our model fails to achieve the face de-identification task, producing the same reference with some variations in facial attributes. This failure is caused by errors when capturing the ground truth image (Gaussian noise), the limited privacy provided by the 15 Zernike polynomials, and an imbalance and low variability in the data collected from 17 individuals (14 males and 3 females). However, as shown in Fig. 6 of the main document, our approach successfully achieves high face de-identification performance when the model is fine-tuned using the privacy images acquired by our camera prototype and the reference from the CelebA-HQ training set.

Finally, Fig. 13 presents qualitative results obtained with

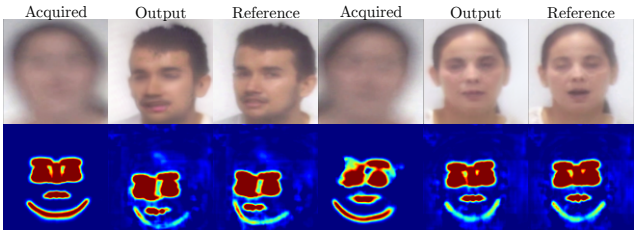

Figure 12. Qualitative results using the ground truth images as reference.

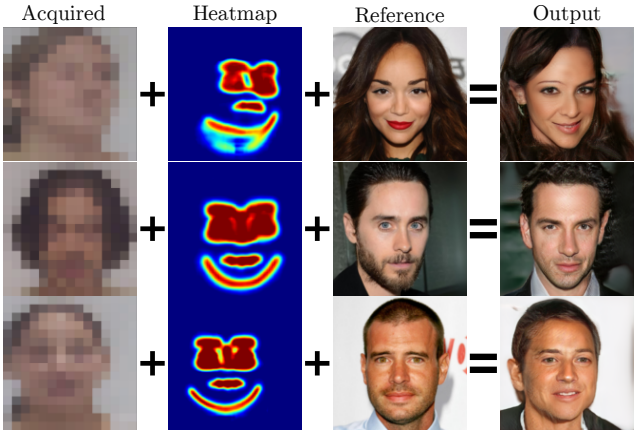

Figure 13. Qualitative results of low-resolution camera, using the data acquired in the Lab.

a low-resolution camera from the lab data. The model was fine-tuned using the same hyperparameters as above and the references from the CelebA-HQ training set. These qualitative results with the quantitative results presented in the main document (Tab. 4), demonstrate the effectiveness of our method when using both, low-resolution and our proposed privacy-preserving camera.

13. Concluding remarks

As demonstrated in Section 8, our method, while novel, shares a vulnerability common to encryption systems: if an attacker gains access to the encryption key (in our case, the Point Spread Function (PSF) of the designed camera) they can use non-blind deconvolution algorithms to potentially recover information that may reveal the identity of the individuals. However, the difficulty of this recovery depends on the level of distortion introduced by the camera. There is an inherent trade-off between the privacy level of acquired images and the performance of face de-identification tasks. Also note that our proposed approach offers the advantage of being both simulatable and implementable as a software tool within the firmware of a camera, eliminating the need for actual lens fabrication. This flexibility enhances the practical feasibility of our method. Finally, our primary contribution is introducing privacy-preserving optics for high-level computer vision tasks, such as face de-identification. However, we showed that using other privacy cameras such as low-resolution can also be integrated into our privacy-preserving face de-identification framework. We believe our work marks a step forward in the development of more secure and practical privacy-preserving vision systems.

14. Personal data/human subjects discussion

In this work, we acquired videos from the faces of persons in our optics Lab. In total, 17 subjects between 20 and 40 years old collaborated to acquire the videos with our prototype camera described in section 4.5 of the main document. We elaborated an informed consent document to explain our research and how we would use their video data. The subjects who accepted participating in the project signed a hard copy of the informed consent document and sent it to us. No approval from the institutional review board (IRB) was required in the country where we acquired the videos.

References

[1] Vivek Boominathan, Jesse K Adams, Jacob T Robinson, and Ashok Veeraraghavan. Phlatcam: Designed phase-mask based thin lensless camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 3, 4

[2] Julie Chang, Vincent Sitzmann, Xiong Dun, Wolfgang Heidrich, and Gordon Wetzstein. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Scientific Reports*, 2018. 1

[3] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[4] Jiangxin Dong, Stefan Roth, and Bernt Schiele. Deep wiener deconvolution: Wiener meets deep learning for image deblurring. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 2

[6] Joseph W Goodman. *Introduction to Fourier optics*. Macmillan Learning, 4 edition, 2017. 1, 2

[7] Carlos Hinojosa, Juan Carlos Niebles, and Henry Arguello. Learning privacy-preserving optics for human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[8] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[10] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 4

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[12] 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)*, 2019. 4

[13] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*. PMLR, 2018. 2

[14] Francesco Pittaluga and Sanjeev J Koppal. Privacy preserving optics for miniature vision sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 4

[15] 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)*, 2018. 1

[16] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *IEEE International Conference on Computer Vision (CVPR)*, 2019. 2

[17] Yasin Yaz, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, Vijay Chandrasekhar, et al. The unusual effectiveness of averaging in gan training. In *International Conference on Learning Representations (ICLR)*, 2018. 2