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/

Abstract

The modern surge in camera usage alongside widespread computer vision technology applications poses significant privacy and security concerns. Current artificial intelligence (AI) technologies aid in recognizing relevant events and assisting in daily tasks in homes, offices, hospitals, etc. The need to access or process personal information for these purposes raises privacy concerns. While software-level solutions like face de-identification provide a good privacy/utility trade-off, they present vulnerabilities to sniffing attacks. In this paper, we propose a hardware-level face de-identification method to solve this vulnerability. Specifically, our approach first learns an optical encoder along with a regression model to obtain a face heatmap while hiding the face identity from the source image. We also propose an anonymization framework that generates a new face using the privacy-preserving image, face heatmap, and a reference face image from a public dataset as input. We validate our approach with extensive simulations and hardware experiments.

1. Introduction

The widespread use of cameras, coupled with the ubiquitous application of computer vision technology across various facets of our daily lives, has led to a significant and increasing concern about privacy and security. Specifically, current AI technology enables systems to recognize relevant events and assist us in daily activities in homes, offices, hospitals, etc. However, how can we ensure this technology protects our privacy, especially when it needs to access or process our personal information? This rising concern over data privacy has led international entities like the European Union to establish regulations for ensuring privacy [43].

Different studies have explored methods and applications for protecting privacy in computer vision. In general, these studies can be divided into software and hardware-level protection. Most methods operate at the software-level since they directly work on already acquired high-fidelity images. Prior approaches rely on domain knowledge and hand-crafted approaches, such as pixelation, blurring, and face/object replacement, to protect sensitive information [24, 36, 52]. However, traditional anonymization methods, such as face blurring, change the image semantics, resulting in a significant detection performance drop in downstream tasks such as face detection [15] and tracking [44]. Recent studies have shown that de-identifying people in images, *i.e*., removing the identification characteristics and generating virtual faces to replace these identities, maintains the utility for downstream tasks [3, 14, 29, 48].

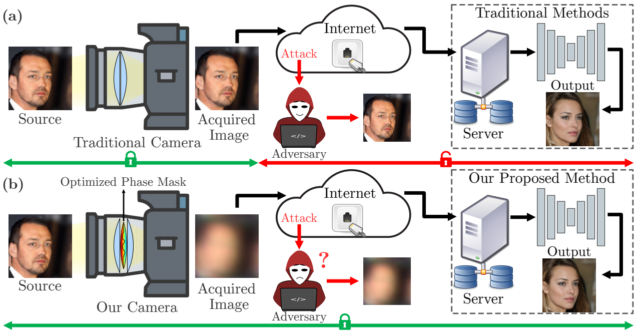

Figure 1. **Traditional *vs*. our proposed pipeline.** (a) The traditional face de-identification pipeline is vulnerable to adversarial attacks (🔓). (b) We propose to close the security gap and enhance protection by learning privacy-preserving optics (🔒).

Unfortunately, these software-level solutions are vulnerable to adversarial attacks that can get direct access to the original privacy-sensitive images before the vision algorithms process them (see Fig. 1 (a)). Consequently, hardware-level privacy protection approaches are considered more secure, since they rely on the optical system to add an extra layer of security by removing sensitive data during image acquisition. Prior hardware-level methods include the use of low-resolution cameras [38], defocus lenses [34], and depth cameras [41] to perform different computer vision tasks like human action recognition and pose estimation. Recent approaches employ the end-to-end and joint optimization of the optics and vision task to promote privacy [12, 42]. With such approaches, it is possible to learn a phase mask [40] to modulate the incident light field to filter out privacy-sensitive information before image capture.

[*] Project lead; † Equal PI contribution.

Paper Contribution. Among the privacy-preserving solutions, face de-identification using generative adversarial networks (GAN) provides the best privacy/utility trade-off. However, current face de-identification methods operate at the software-level, introducing a security gap between image acquisition and algorithms (see Fig. 1 (a)). In this paper, we develop a hardware-level face de-identification approach to close the security gap as shown in Fig. 1 (b). In our study, we observe that the learned optics filter out high-frequency information from the image to protect privacy and promote face misrecognition; however, low-frequency information is preserved. Therefore, we first perform end-to-end training of our privacy-preserving computational camera (a.k.a optical encoder) and a heatmap regression model to obtain a face heatmap that conceals privacy-sensitive information from the source image. Then, we train a GAN using the acquired privacy-preserving image, the corresponding face heatmap, and a reference face image from a public dataset to generate a new synthetic face. The acquired privacy-preserving image and the face heatmap conceal the true identity of the source but provide general information that allows our model to understand the global geometry (e.g., head position). On the other hand, the given reference image is used to extract the specific style for the new face. We summarize our contributions as follows: **(i)** We introduce a full privacy-preserving framework for face de-identification that addresses the security gap between image acquisition and algorithms in traditional approaches. **(ii)** Our approach consists of two integrated stages: First, we jointly optimize the lens of a computational camera to encode the source images such that the true identity of the person is concealed, but important features are preserved to perform face heatmap regression. Second, we use the acquired privacy-preserving images, the corresponding face heatmaps, and reference images from a public dataset to train a GAN to generate a new face. **(iii)** We propose a facial expression consistency loss and use an LPIPS-based loss term to improve quality and promote the preservation of detailed expressions (like smiles) in the generated images. **(iv)** We perform extensive simulations on FFHQ and CelebA-HQ datasets to validate our proposed approach. Furthermore, we build a prototype camera and perform hardware experiments that match simulations.

2. Related Work

We categorize previous efforts in privacy-preserving vision into two groups: *software-level* and *hardware-level* protection, where the latter is considered more robust to attacks.

2.1. Software-level Privacy Protection

In the literature, most privacy-preserving vision approaches work at the software level. Specifically, such methods only perform software-level processing on high-resolution images acquired by traditional cameras.

Non-generative Approaches. Traditional methods leverage domain knowledge and use hand-crafted techniques such as blurring, mosaicing, masking, pixelation, and face/object replacement to protect sensitive information [1, 21, 24, 33, 36, 52]. These methods can successfully conceal the identity of a person and private sensitive objects on an image when the target to protect is known beforehand. However, they also could introduce artifacts harming the performance of downstream visual tasks [15].

GAN-based Face De-identification. On the other hand, face de-identification using GAN provides a better privacy/utility trade-off than other approaches. These methods focus on generating virtual faces to replace the original identity, preserving privacy and avoiding being recognized by unauthorized users (adversaries/hackers) and computer vision systems. Several methods in the literature [3, 6, 14, 23, 27, 45, 48] are built on top of state-of-the-art GAN networks like StyleGAN2 [18] and StarGAN2 [8] given their impressive ability to capture the distribution of the training samples and then generate good-looking face images. Among prior works, DeepPrivacy [14] first extracts the sparse facial key points from the image and masks the face region with a constant value; then employs a StyleGAN2 generator to in-paint a randomly generated face, preserving the contextual and pose information. Similarly, CIAGAN [29] also masks the source face region and uses a conditional GAN to perform conditional identity swapping by employing an identity discriminator to force the generator to synthesize a new face on the masked region. Most recent approaches aim to improve the ability of the models to generate better quality and natural-looking faces while keeping similar privacy preservation results as CIAGAN and DeepPrivacy [3, 23, 27, 48]. For example, Barattin *et al.* [3] proposes to directly optimize the latent space to ensure the identity is distant enough from the original image while preserving some facial attributes. While these software-level approaches preserve privacy in the final application, the acquired images are not protected. This security gap makes these methods vulnerable to adversarial attacks that can get direct access to the original privacy-sensitive image captures before the algorithms process them.

2.2. Hardware-level Privacy Protection

These approaches are considered more secure since they rely on the optical system to add an extra layer of security that removes sensitive data during image acquisition.

Fixed Optics. Prior methods include the use of low-resolution cameras to capture images and videos, avoiding the unwanted leak of identity information from human subjects [37, 38]. Similarly, authors in [34, 35] propose optical designs based on a defocusing lens to filter or block sensitive information directly from the incident light field before sensor measurement acquisition and perform K-anonymity to protect privacy. In particular, they show how to select a defocus blur that provides a certain level of privacy over a working region within the sensor size limits; however, only using optical defocus for privacy may be susceptible to reverse engineering attacks [12, 13].

Learning Optics. Instead of using fixed optics privacy-preserving cameras to acquire the data and then train algorithms, the privacy/utility trade-off can be improved at the hardware-level by approaches that jointly optimize the optics and vision algorithms to directly filter sensitive data and enhance utility in downstream tasks [12, 40]. Most recent approaches learn a phase mask [12, 46] to modulate the incident light field and filter out privacy-sensitive information before image acquisition. This idea has been applied in different tasks, such as human pose estimation [12], human action recognition [13], image captioning [2], and passive depth estimation [42]. Our proposed framework, developed in the next section, uses privacy-preserving cameras to improve privacy and close the security gap while preserving the utility given by face de-identification methods. To the best of our knowledge, this is the first work that learns optics for the face de-identification task. Note that our framework can be used with fixed optics cameras, e.g., low-resolution cameras; however, learning the optics provides better control over privacy and utility, as we show in Section 4.

3. Proposed Methodology

We are interested in using privacy-preserving images to add an extra layer of security to the face de-identification task. Previous approaches [2, 12, 13] learn optics for a specific task, and the utility of the acquired privacy-preserving images is limited only to this task. To overcome this limitation, we propose to combine optics learning with a generative model to maximize utility in downstream tasks. Specifically, the output of our framework is a realistic face image that maintains the anonymity of the original person.

In general, our framework (Fig. 2) consists of two parts: optics learning and new face generation. In the first part, our strategy is to jointly optimize the camera optics and a heatmap regression network to obtain a face heatmap while ensuring privacy protection. Our key observation is that we can learn the camera lens to degrade the image such that the high-frequency information is filtered out, concealing the subject identity while preserving important features to obtain a face heatmap. We can then leverage the face heatmap to infer the global geometry information of the original (source) image like head pose and eyes, nose, and mouth position. In the second part, our main goal is to learn the parameters of a conditional GAN to generate a new face identity from the input privacy-preserving image, the face heatmap, and a reference image.

3.1. Optics Learning

The main goal of the optical component in our proposed approach (Fig. 2) is to design a phase mask to visually distort images by removing the high-frequency spatial information and hence conceal privacy-sensitive attributes while preserving useful information to extract a face heatmap.

Image Formation Model and Phase Mask Learning. We adopt a similar strategy as previous works in [12, 13, 40] to couple the modeling and design of two essential operators in the imaging system: wave propagation and phase modulation. We model the image acquisition process by defining the point spread function (PSF) in terms of the lens surface profile. This allows emulating the propagation of the wavefront and training the parameters of the refractive lens. Considering the Fresnel approximation and paraxial regime [10] for incoherent illumination, we describe the PSF as:

$$H(u',v') = |\mathcal{F}^{-1}\{\mathcal{F}\{P(u,v)\cdot W(u,v)\}\cdot T(f_u,f_v)\}|^2, \quad (1)$$

where $P(u,v) = P_t(u,v)\cdot P_\phi(u,v)$, $W(u,v)$ is the incoming wavefront, $T(\cdot)$ represents the transfer function with (f_u,f_v) as the spatial frequencies, $P_\phi(u,v) = \exp(-ik\phi(u,v))$ with $\phi(u,v)$ as the phase mask and $k = 2\pi/\lambda$ as the wavenumber, $P_t(u,v) = \exp\left(-i\frac{k}{2d}(u^2+v^2)\right)$ denotes the light wave propagation phase with d as the object-lens distance, $\mathcal{F}\{\cdot\}$ denotes the 2D Fourier transform, and (u',v') is the spatial coordinate on the camera plane. The values of $\phi(\cdot)$ are modeled via Zernike polynomials with $\phi(u,v) = \mathcal{R}(\sqrt{u^2+v^2})\cdot\cos\left(\arctan\left(v/u\right)\right)$, where $\mathcal{R}\left(\cdot\right)$ represents the radial polynomial function [22]. To train the phase mask values using our model, we discretize the phase mask $\phi(\cdot)$ as:

$$\boldsymbol{\phi} = \sum_{j=1}^{p}\beta_j\mathbf{Q}_j, \quad (2)$$

where $\mathbf{Q}_j$ denotes the j-th Zernike polynomial in Noll notation, and β_j is the corresponding coefficient [4]. Each Zernike polynomial describes a wavefront aberration [22]; hence the phase mask $\boldsymbol{\phi}$ is formed by the linear combination of p aberrations. In this regard, the optical element parameterized by $\boldsymbol{\phi}$ can be seen as an optical encoder, where the coefficients $\boldsymbol{\beta} = \{\beta_j\}_{j=1}^{p}$ determine the data transformation. Therefore, our end-to-end training finds the coefficients $\boldsymbol{\beta}$ that provide the maximum visual distortion of the scene but allow extracting relevant features to obtain a face heatmap via a regression network.

The sensing process of our camera can be modeled as a shift-invariant convolution operation between the source image (scene) and the PSF as:

$$\mathbf{y} = \mathcal{C}\left(\mathbf{H}*\mathbf{x}\right)+\boldsymbol{\eta}, \quad (3)$$

where $\mathbf{x}$ represents the discrete color source image of the person we want to protect; $\mathbf{H}$ denotes the discretized version of the PSF [10] in Eq. (1); $\boldsymbol{\eta}$ represents Gaussian noise in the sensor, and $\mathcal{C}(\cdot)$ is the camera response function, which we assume linear. Please refer to our supplementary for more details on our light propagation model.

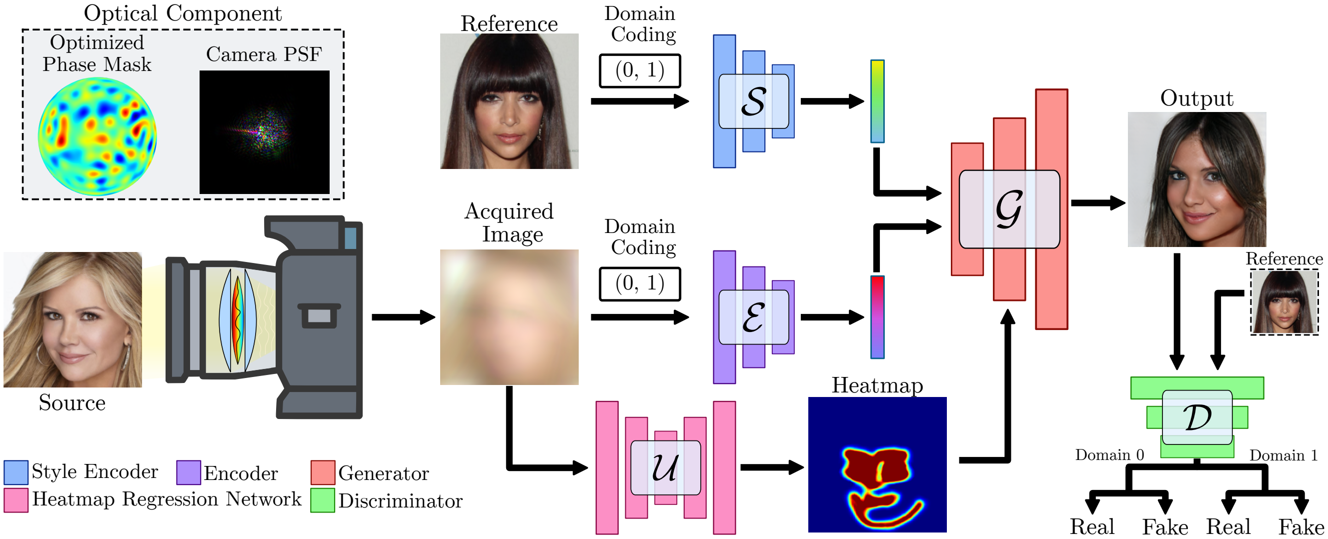

Figure 2. **Our proposed face de-identification framework**. We first jointly optimize the camera optics and a heatmap regression network to obtain a face heatmap and conceal source image identity. Then, we train a GAN network to generate a new face identity using the acquired image, the obtained heatmap, and a reference image. We leverage global geometry information (e.g. face pose and eyes, nose, mouth position) from the acquired image and the heatmap while the style is extracted from the reference image.

Heatmap Regression Network. We adopt the same implementation as in [8], which is an adaptation from the Face Alignment Network (FAN) [5] and the adaptive wing-based heatmap [47]. This network consists of stacked Hourglass (HG) [32] layers and coordinate convolutions (CoordConv) [26], which encode coordinate information as additional channels before the convolution operation. Given an image $\mathbf{y}$ captured by our camera, we use the regression network $\mathcal{U}$ to obtain a heatmap $\mathbf{m} = \mathcal{U}(\mathbf{y})$ as shown in Fig. 2. This heatmap fits the eyes, nose, mouth, and chin region and retrieves geometry information about the face.

3.2. Generative Network

To perform the face image generation, we use the generator, style encoder, and discriminator network architectures from StarGAN2 [8]. Considering $\omega \in \Omega$ as an arbitrary domain and given a privacy-preserving image $\mathbf{y}$ captured by our camera as in Eq. 3, we first use an inversion encoder $\mathcal{E}$ to map $\mathbf{y}$ to the latent space $\bar{\mathbf{y}} = \mathcal{E}(\mathbf{y})$. Then, we want the generator $\mathcal{G}$ to translate $\mathbf{y}$ into a new image $\mathcal{G}(\bar{\mathbf{y}}, \mathbf{s}, \mathbf{m})$ reflecting a domain-specific style code $\mathbf{s}$ provided by the style encoder $\mathcal{S}$ and the pose and face geometry information provided by the heatmap $\mathbf{m}$. We extract the style code $\mathbf{s} = \mathcal{S}_\omega(\mathbf{r})$ from a publicly available reference image $\mathbf{r}$ corresponding to a specific domain ω. Note that using different reference images produces different style codes $\mathbf{s}$. Specifically, $\mathbf{s}$ represents the style of a particular reference image $\mathbf{r}$ and domain ω; hence it is not necessary to directly provide ω to the generator $\mathcal{G}$ and allows $\mathcal{G}$ to synthesize images of all domains reflecting the style of $\mathbf{r}$. Finally, the discriminator $\mathcal{D}$ is a multi-task discriminator [25, 31], which consists of multiple output branches. Each branch $\mathcal{D}_\omega$ learns a binary classification determining whether an image $\mathbf{r}$ is a real image of its domain ω or a fake image $\mathcal{G}(\bar{\mathbf{y}}, \mathbf{s}, \mathbf{m})$ produced by $\mathcal{G}$. Note that, unlike StarGAN2, we use the reference image $\mathbf{r}$ in the discriminator instead of the source image $\mathbf{x}$ to avoid privacy information leakage.

3.3. Training Objectives

We divide our training into two stages. In the first stage, we learn the coefficients of Zernike polynomials $\boldsymbol{\beta}$ and the parameters of the heatmap regression network to produce images that visually conceal the identity of the person while allowing to extract a face heatmap. For the second stage, we freeze the camera and regression model and only train the generative network to translate the acquired privacy-preserving image to a new face using the extracted heatmap and a reference face image.

3.3.1 Stage I: Optical Design

This stage aims to extract the face heatmap information from $\mathbf{y}$ by learning the optics. As the starting point of our training, we assume an aberration-free lens and use a pre-trained regression model $\mathcal{U}$.

Optics Loss Function. To encourage image degradation, we minimize the difference of the acquired image $\mathbf{y}$ with the original image $\mathbf{x}$. Also, for easy and faster convergence, we promote symmetric and low frequencies using a predefined defocus PSF $\mathbf{H}_f$ as a regularizer. Specifically, assuming we acquire $\mathbf{y}$ using Eq. 3 and the images are normalized between $[0, 1]$, we define the loss function for our camera lens optimization as:

$$\mathcal{L}_{optics} = \mathbb{E}_{\mathbf{x}} \left[1 - \|\mathbf{x} - \mathbf{y}\|_2^2 + \alpha_1 \|\mathbf{H} - \mathbf{H}_f\|_2 \right], \quad (4)$$

where $\| \cdot \|_2^2$ is the standard mean squared error (MSE), and α_1 is a hyperparameter. Note that $\mathbf{H}_f$ is a predefined

PSF whose parameters are frozen during training. See our supplementary document for details on our regularizer $\mathbf{H}_f$.

Heatmap Regression Loss Function. After initializing $\mathcal{U}$ with the pre-trained weights [47], we aim to obtain a heatmap from $\mathbf{y}$ that closely resembles the heatmap obtained from $\mathbf{x}$; therefore, we use the following loss:

$$\mathcal{L}_{hmap} = \mathbb{E}_{\mathbf{x}} \left[\| \mathcal{U}(\mathbf{y}) - \mathcal{U}^*(\mathbf{x}) \|_1 \right], \tag{5}$$

where we use the heatmap $\mathcal{U}^*(\mathbf{x})$ as ground truth and $\mathcal{U}^*$ is a freeze pre-trained regression model.

Full Objective. We finetune the camera model and the regression network $\mathcal{U}$ end-to-end to gradually distort the optics and learn to extract the heatmap from the $\mathbf{y}$ using the following full objective function:

$$\min_{\mathcal{C},\mathcal{U}} \mathcal{L}_{optics} + \alpha_2 \mathcal{L}_{hmap}, \tag{6}$$

where α_2 is a hyperparameter to control the similarity between $\mathbf{H}$ and $\mathbf{H}_f$.

3.3.2 Stage II: Generative Network

Once we train the lens and the regression network parameters, we now train our generative network using the following objectives. Note that the first four losses are adopted from StarGAN2 [8], while the last two losses are specifically proposed to improve the generation when using our privacy-preserving image $\mathbf{y}$.

Adversarial Objective. During training, we randomly select a reference image $\mathbf{r}$ and a target domain $\tilde{\omega} \in \Omega$, from which we compute a target style code $\tilde{\mathbf{s}} = \mathcal{S}_{\tilde{\omega}}(\mathbf{r})$. The generator $\mathcal{G}$ receives the latent code $\bar{\mathbf{y}} = \mathcal{E}(\mathbf{y})$ and the mask $\mathbf{m} = \mathcal{U}(\mathbf{y})$ extracted from the image $\mathbf{y}$ and the style code $\tilde{\mathbf{s}}$ as inputs and is trained to produce an output image $\mathcal{G}(\bar{\mathbf{y}}, \mathbf{m}, \tilde{\mathbf{s}})$, guided by an adversarial loss function:

$$\begin{aligned} \mathcal{L}_{adv} = &\mathbb{E}_{\mathbf{y},\omega} \left[\log \mathcal{D}_{\omega}(\mathbf{r}) \right] \\ &+ \mathbb{E}_{\mathbf{y},\tilde{\omega},\mathbf{r}} \left[\log(1 - \mathcal{D}_{\tilde{\omega}}(\mathcal{G}(\bar{\mathbf{y}}, \mathbf{m}, \tilde{\mathbf{s}}))) \right], \end{aligned} \tag{7}$$

where $\mathcal{D}_{\omega}(\cdot)$ denotes the output of the discriminator $\mathcal{D}$ corresponding to the domain ω.

Style Reconstruction. To ensure that the generator $\mathcal{G}$ effectively incorporates the style code $\tilde{\mathbf{s}}$ in the image generation process, we employ a style reconstruction loss:

$$\mathcal{L}_{sty} = \mathbb{E}_{\mathbf{y},\tilde{\omega},\mathbf{r}} \left[\| \tilde{\mathbf{s}} - \mathcal{S}_{\tilde{\omega}}(\mathcal{G}(\bar{\mathbf{y}}, \mathbf{m}, \tilde{\mathbf{s}})) \|_1 \right]. \tag{8}$$

Style Diversification. To further enable the generator to produce diverse images, we explicitly regularize $\mathcal{G}$ with the diversity sensitive loss [28]:

$$\mathcal{L}_{ds} = \mathbb{E}_{\mathbf{y},\tilde{\omega},\mathbf{r}_1,\mathbf{r}_2} \left[\| \mathcal{G}(\bar{\mathbf{y}}, \mathbf{m}, \tilde{\mathbf{s}}_1) - \mathcal{G}(\bar{\mathbf{y}}, \mathbf{m}, \tilde{\mathbf{s}}_2) \|_1 \right], \tag{9}$$

where the target style codes $\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$ are produced by $\mathcal{S}$ using two reference images $\mathbf{r}_1$ and $\mathbf{r}_2$.

Preserving Source Characteristics. To ensure that the domain-invariant attributes (such as pose) of the input source image $\mathbf{x}$ are accurately maintained in the generated image $G(\bar{\mathbf{y}}, \mathbf{m}, \tilde{\mathbf{s}})$, we utilize the cycle consistency loss that leverages the information from the mask extracted from the privacy-preserving image $\mathbf{y}$.

$$\mathcal{L}_{cyc} = \mathbb{E}_{\mathbf{y},\omega,\tilde{\omega},\mathbf{r}} \left[\| \mathbf{y} - \mathcal{G}(\hat{\mathbf{y}}, \hat{\mathbf{m}}, \hat{\mathbf{s}}) \|_1 \right], \tag{10}$$

where $\hat{\mathbf{y}} = \mathcal{E}(\mathcal{G}(\bar{\mathbf{y}}, \mathbf{m}, \tilde{\mathbf{s}}))$, $\hat{\mathbf{m}} = \mathcal{U}^*(\mathcal{G}(\bar{\mathbf{y}}, \mathbf{m}, \tilde{\mathbf{s}}))$, $\hat{\mathbf{s}} = \mathcal{S}_{\omega}(\mathbf{y})$ is the estimated style code of the input privacy-preserving image $\mathbf{y}$, and ω corresponds to the original domain of $\mathbf{y}$, which is the same as the source image $\mathbf{x}$.

LPIPS. We employ an LPIPS loss term [51] to improve the generated image quality and more effectively reflect the reference image face attributes.

$$\mathcal{L}_{LPIPS} = \mathbb{E}_{\mathbf{y},\tilde{\omega},\mathbf{r}} \left[\| \mathcal{Q}(\mathbf{r}) - \mathcal{Q}(\mathcal{G}(\bar{\mathbf{y}}, \mathbf{m}, \tilde{\mathbf{s}})) \|_2 \right], \tag{11}$$

where $\mathcal{Q}$ is the pre-trained perceptual feature extractor.

Facial Expression Consistency. Since the high-frequency information is missing in the privacy-preserving image $\mathbf{y}$, it is challenging to preserve small expressions like smiles, eyes blinking, or small mouth movements in the generated image. Therefore, we propose a facial expression consistency loss as follows:

$$\mathcal{L}_{expr} = \mathbb{E}_{\mathbf{x}} \left[\| \mathbf{m}^* \odot \mathbf{x} - \mathbf{m} \odot \mathbf{y} \|_1 \right], \tag{12}$$

where $\odot$ denotes the element-wise product, and $\mathbf{m}^* = \mathcal{U}^*(\mathbf{x})$ is the heatmap obtained from the source image $\mathbf{x}$.

Full Objective. Our full objective function can be summarized as follows:

$$\begin{aligned} \min_{\mathcal{G},\mathcal{S},\mathcal{E}} \max_{\mathcal{D}} \Big(&\mathcal{L}_{adv} + \lambda_{sty}\mathcal{L}_{sty} - \lambda_{ds}\mathcal{L}_{ds} + \lambda_{cyc}\mathcal{L}_{cyc} \\ &+ \lambda_{LPIPS}\mathcal{L}_{LPIPS} + \lambda_{expr}\mathcal{L}_{expr} \Big), \end{aligned} \tag{13}$$

where $\lambda_{sty}, \lambda_{ds}$, λ_{cyc}, λ_{LPIPS}, and λ_{expr} are hyperparameters for each term. See our supplementary document for training details of Stage I and Stage II, and results when using latent vectors instead of reference images, as used in StarGAN2, to generate the style codes.

4. Experimental Results

We extensively evaluate the performance of our proposed face de-identification approach using standard metrics and evaluation protocols and compare it with other methods.

Dataset. We train our proposed model on two widely used datasets: CelebA-HQ [16] and FFHQ [17]. We separate the images of the CelebA-HQ dataset into two domains (male and female) as in [8]. We show our main results and ablation studies with the CelebA-HQ dataset using the same train/eval set split provided by [8]. To quantitatively compare with other methods, we follow the common practice of using the FFQH dataset for training and the CelebA-HQ dataset for evaluation [23, 48]. For a fair comparison, all images are resized to 256 × 256 resolution for training, which is the same resolution used in the baselines.

Metrics. Following previous works [6, 48], we quantitatively evaluate the identity protection given by our face de-identification approach by measuring the ℓ_2 distance (DIS) of embedding vectors from the de-identified and original faces extracted by a pre-trained face recognition model. Specifically, we employ the Face Recognition library (FR) [9] and FaceNet [39], which is pre-trained on two public datasets: CASIA-Webface [49] and VGGFace2

Components			DIS↑			Landmarks ↓		Bounding Box ↓		FID↓
H	L	E	FR	CASIA	VGGFace2	MtCNN	Dlib	MtCNN	Dlib	
×	×	×	**0.851**	1.132	1.313	5.434	6.473	13.667	8.008	34.388
✓	×	×	0.847	1.162	1.350	5.309	6.260	12.988	7.654	36.056
✓	✓	×	0.847	**1.174**	**1.359**	4.692	5.683	11.818	7.037	29.338
✓	×	✓	0.664	1.037	1.199	**1.884**	**2.081**	**4.694**	**4.365**	24.998
✓	✓	✓	0.697	1.073	1.255	2.022	2.259	4.727	4.985	**24.972**

Table 1. **Quantitative results of the ablation studies**. The best result for each pre-trained face recognition model is in bold, and the second-best result is underlined.

[7]. We further evaluate the utility (↑) of our generated images and the preservation of facial attributes from the source image (↓), such as face orientation, lips position, and other attributes, using the Dlib library [20] and MtCNN network [50]. Specifically, we estimate the facial landmarks between the original and de-identified faces and measure their similarity using the ℓ_1 norm. Additionally, we estimate the facial bounding box to evaluate whether the face is located in the correct position and size. Finally, we use Fréchet inception distance (FID ↓) to measure the distance between the generated and real distribution. Since our proposed approach uses reference-guided synthesis, each source image is transformed using 10 reference images randomly sampled from the test set and we report the average value.

4.1. Ablation Studies

We conducted a series of experiments to investigate the contribution of using the face heatmap (H), $\mathcal{L}_{LPIPS}$ (L), and $\mathcal{L}_{expr}$ (E) loss functions on the performance of our method. Table 1 shows the results of our ablation studies. In the table, a checkmark (✓) denotes the inclusion, and a red cross (×) denotes the exclusion of specific components. From the table, we observe that there is an improvement in all metrics when face heatmaps (H ✓) are used. This supports our hypothesis that integrating a face heatmap in our framework effectively maintains geometric information of the face (such as head pose) from the source image. Since we obtain the face heatmap directly from the privacy-preserving image, the source image privacy is still preserved. Similarly, we observe that the inclusion of the $\mathcal{L}_{LPIPS}$ (L ✓) and $\mathcal{L}_{expr}$ (E ✓) significantly improves all the metrics, especially the FID and facial Landmarks, demonstrating their contribution to our framework to preserve detailed expressions (like smiles) in the generated images. Note that we input the reference image instead of the source image to our discriminator. We conducted experiments using the source image instead of the reference in the discriminator. The results show an improvement in the FID; however, the face anonymization metrics (DIS) decrease significantly, which suggests that the face information from the source image is leaked to the generator. Hence, we only use the reference image in our discriminator to maximize privacy preservation. See our supplementary document for results when using the source image in the discriminator.

Method	DeepPrivacy	CIAGAN	FIT	IDeudemon	RiDDLE	Ours-LR	Ours
FR ↑	0.783	0.671	0.812	**0.819**	0.776	0.795	0.734
CASIA ↑	1.091	0.919	1.207	**1.228**	1.033	1.123	1.071
VGGFace2 ↑	1.187	1.085	1.224	1.233	1.129	**1.294**	1.251

Table 2. **Quantitative Results - DIS**. We compare our method against SOTA face de-identification methods on the CelebA-HQ dataset using the DIS metrics (FR, CASIA, VGGFace2).

Method		CIAGAN	FIT	DeepPrivacy	RiDDLE	Ours-LR	Ours
FID ↓		32.611	30.331	23.713	**15.389**	34.161	29.218
FD ↑	MtCNN	0.992	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	Dlib	0.937	0.984	0.980	0.991	**0.994**	0.991
BB ↓	MtCNN	20.387	7.879	4.654	**3. 824**	5.358	5.314
	Dlib	15.476	4.218	2.685	**1.700**	4.694	4.821
LM ↓	MtCNN	8.042	3.572	3.280	**1.674**	2.392	2.298
	Dlib	8.930	4.047	2.896	**1.512**	2.749	2.716

Table 3. **Quantitative Results - Detection.** We compare our method against SOTA face de-identification methods on the CelebA-HQ dataset using the Face Detection (FD), Bounding Box (BB), and Landmarks (LM) metrics.

4.2. Comparison with other methods

Quantitative Results. We quantitatively compare our framework with several state-of-the-art (SOTA) face de-identification approaches, including: CIAGAN [30], Face Identity Transformer (FIT) [11], DeepPrivacy [14], IDeudemon [48], and RiDDLE [23]. We also compare the performance of our proposed framework when using a low-resolution (LR) camera instead of learning the lens. Specifically, we simulate an LR camera (fixed optics parameters) and perform the two-stage training described in section 3 without using the Eq. 4. Table 2 reports the results for all the methods evaluated on the CelebA-HQ dataset using the ℓ_2 distance (DIS) metrics. As observed, our approach achieved the best performance in the VGGFace2 metric and a similar performance in the other metrics. Also, Ours approach using a low-resolution camera (Ours-LR) performs better than our learned lens (Ours); however, LR is susceptible to reverse engineering attacks as shown in section 4.3.

Additionally, we also compare our method against SOTA using face detection (FD), bounding box detection (BB), and landmark (LM) metrics in Table 3. The results show that our approach does not outperform some SOTA methods, such as RiDDLE. The main reason is the significant loss of high-frequency information on the acquired privacy-preserving images, which makes it challenging to generate images that preserve detailed expressions and face geometry from the source image. Although including our proposed loss functions and using heatmaps helps mitigate this problem, a trade-off exists between achieved image degradation and the generative capabilities of our framework. However, even with such loss of information, our method outperforms other methods, such as CIAGAN and FIT.

Qualitative Results. Figure 3 shows generated faces using our proposed framework. In the first row, we show the source image, which is the input of our privacy-preserving

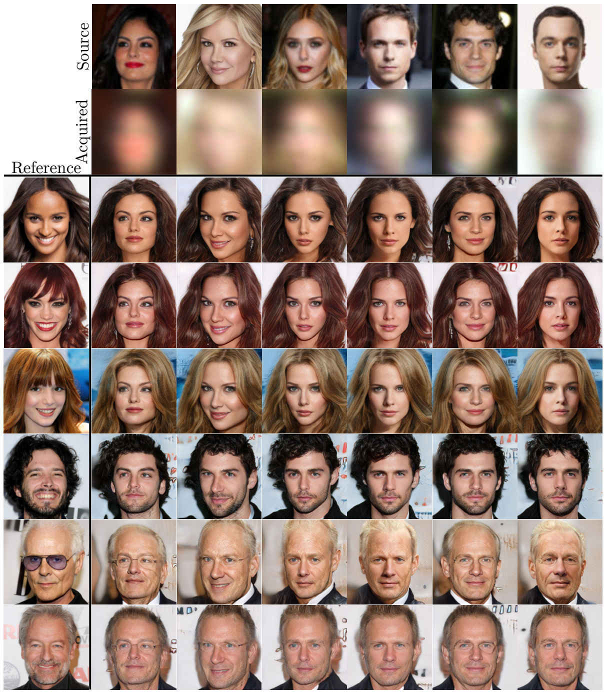

Figure 3. **Qualitative results**. We qualitatively evaluate our proposed face de-identification method using the CelebA-HQ dataset.

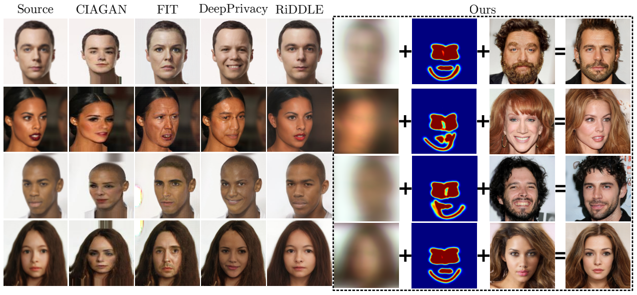

Figure 4. **Qualitative Results**. We compare our method against SOTA methods: CIAGAN, FIT, DeepPrivacy, and RIDDLE.

camera. The image acquired by our camera is shown in the second row. Our generative network generates new faces using the acquired privacy-preserving image and different reference images shown in the first column. We observe that the generated images have the style extracted from the reference images and face geometry of the source images. Furthermore, we qualitatively compare our approach with SOTA methods in Fig. 4 using the CelebA-HQ dataset. While our generated faces might not achieve the same level of realism as those produced by methods like RiDDLE, they effectively fulfill our primary goal of preserving privacy.

4.3. Robustness to Deconvolution

Assuming that an attacker performs an adversarial attack and successfully captures the privacy-preserving image (see Fig. 1 (b)). Then, the attacker could use a deconvolution method to recover the identity of the person. To test the robustness of our optimized lens to deconvolution attacks, we use a SOTA diffusion model (DDRM) proposed by Kawar et al. [19] to recover the original image. DDRM is trained to recover images in multiple scenarios, such as superresolution, inpainting, colorization, and deblurring. We apply the pre-trained DDRM model on privacy-preserving images captured by our optimized lens and a low-resolution camera with 16×16 sensor size. As shown in Fig. 5, DDRM can recover the identity of the person from the low-resolution with high accuracy but fails to recover the face from the privacy-preserving image acquired by our optimized lens. Similar to previous works [12, 13], we also investigate the case where an attacker has direct access to the camera; hence, it can get the PSF by imaging a point of source light and perform the attack using non-blind deconvolution methods (*e.g.*, the Wiener Filter). See our supplementary document for results on non-blind deconvolution attacks.

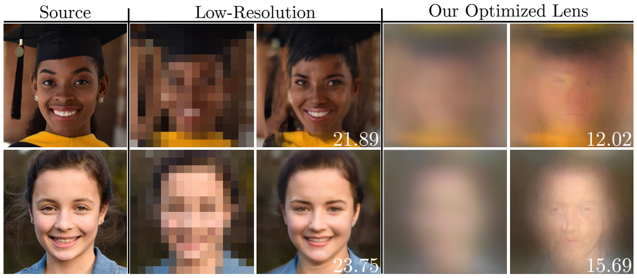

Figure 5. **Robustness to Deconvolution.** We use the SOTA diffusion model in [19] to recover the person identity. We show the PSNR between the original and recovered face in each image.

4.4. Human Evaluation for Privacy Protection

We conduct a human evaluation to assess the effectiveness of privacy protection given by our approach under four different scenarios. In the first scenario, we randomly select five original face images alongside their corresponding "low-resolution" and "blurred" image versions. Here, the term "blurred" refers to the privacy-preserving images acquired through our optimized lens; then, we display the images to 106 individuals and ask them to judge to what degree they believe that the identity of the person is protected considering the following options: *Not Protected*, *Slightly Protected*, *Moderately Protected*, and *Highly Protected*. We found that only 12.76% of the surveyed people consider that the "low-resolution" face image is *Highly Protected* while 56.00% consider that the "blurred" face image is *Highly Protected*. In the second scenario, we generate a new face image with our 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 106 individuals to identify the source image. The responses show that 91.82% of the people fail to identify the source image, showing the

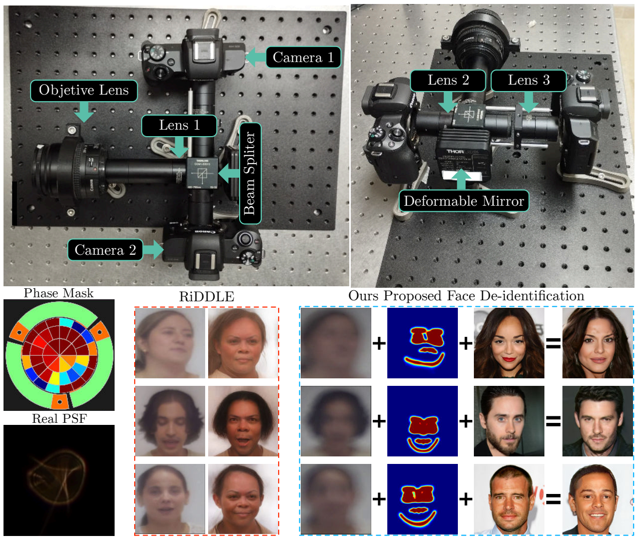

Figure 6. **Proof-of-concept optical system**. (Top) Experimental hardware setup for our privacy-preserving approach. (Bottom) Qualitative results on images acquired by our prototype camera and comparison with the RiDDLE method.

effectiveness of our method. Refer to our supplementary document for details about our human evaluation study.

4.5. Hardware Experiments

We developed a real camera prototype in the optical laboratory to validate our proposed approach; see Fig. 6 (top). Our prototype includes an additional side information branch to obtain the original images without privacy protection, which are used as ground truth to estimate DIS metrics. We emulate the lens designed with our framework using a deformable Thorlabs mirror (DMH40-P01) in a 4f optical system built with 50mm achromatic lenses and two Canon EOS M50 mirrorless cameras as sensors. Due to the deformable mirror limitations, we are restricted to using only $p = 15$ Zernike coefficients. Therefore, considering this limitation, we first simulate the first stage in our proposed framework (Sec. 3.3.1) to obtain the simulated coefficients. Then, we load the learned coefficients to the deformable mirror and calibrate the optical setup. After calibration, we obtained the following learned Zernike coefficients: $\{\beta_1, \beta_2, \beta_3 = 0.0, \beta_4 = -0.83, \beta_5 = 0.0, \beta_6 = -0.31, \beta_7 = 0.08, \beta_8 = -0.69, \beta_9 = 0.0, \beta_{10} = 0.0, \beta_{11} = 0.0, \beta_{12} = -0.67, \beta_{13} = 0.0, \beta_{14} = 0.18, \beta_{15} = 0.0\}$. We acquired 17 short face videos, each lasting between 15 to 30 seconds, to capture the same person with different facial expressions and head positions. After preprocessing, we obtain 3700 face images, which we split into two sets of 3000 and 700 images for finetuning and testing, respectively. Specifically, we finetune the Heatmap Regression Network $\mathcal{U}$ and the Generator $\mathcal{G}$ of our model to adapt them to our acquired dataset. To evaluate the effectiveness of our proposed method on the data obtained in our lab, we use the DIS metric to realize a comparison with the SOTA RiDDLE model [23] over the collected testing set, see Tab. 4. Furthermore, we also show some qualitative comparisons in Fig. 6 (bottom). These visual and quantitative results demonstrate that our low-resolution approach and proposed learned camera are effective for face anonymization in real-world scenarios. Please refer to our supplementary material for a qualitative comparison when using a low-resolution camera with data acquired in our lab.

Method	DIS↑		
	FR	CASIA	VGGFace2
RiDDLE	0.798	0.958	1.224
Ours-LR	**0.808**	1.141	1.310
Ours	0.792	**1.156**	**1.334**

Table 4. **Quantitative results - Prototype camera.** We quantitatively evaluate the performance of our proposed approach using the privacy-preserving data acquired by our prototype camera and compare it against RiDDLE, which uses traditional images.

5. Conclusion

We presented a novel hardware-level face de-identification approach that proposes a solution to close the security gap between image acquisition and algorithms in traditional software-level methods. To ensure training stability and efficiency, we propose a two-stage optimization process. First, we leverage optics learning to capture privacy-preserving images while preserving facial features for face heatmap regression. Then, we employ a generative adversarial network to generate new synthetic faces that preserve facial expression from the source image but style and face appearance from a reference image with no privacy concern. Specifically, our proposed approach offers two key advantages over existing methods as **Hardware-Level Security** mitigating the vulnerability to sniffing attacks that can compromise privacy protection, and **High-Quality Face Generation** using our proposed generative framework for privacy-preserving images acquired with our camera.

Limitations. In our prototype camera, the deformable mirror is the main limitation, which only uses 15 Zernike Polynomials, limiting the scene's level of distortion. Additionally, generating faces using only our privacy-preserving images is still challenging for the generative model; hence, some artifacts could appear in the generated faces.

Acknowledgment. This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding, the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI), and the Universidad Industrial de Santander (UIS), Colombia under the research projects VIE-3735 and VIE-3924.

References

[1] Shafiq Ahmad, Pietro Morerio, and Alessio Del Bue. Person re-identification without identification via event anonymization. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[2] Paula Arguello, Jhon Lopez, Carlos Hinojosa, and Henry Arguello. Optics lens design for privacy-preserving scene captioning. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022. 3

[3] Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Attribute-preserving face dataset anonymization via latent code optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

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

[5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 4

[6] Jingyi Cao, Bo Liu, Yunqian Wen, Rong Xie, and Li Song. Personalized and invertible face de-identification by disentangled identity information manipulation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 5

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE international conference on automatic face & gesture recognition*. IEEE, 2018. 6

[8] 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, 5

[9] Adam Geitgey. Face recognition. `https://github.com/ageitgey/face_recognition`, 2018. 5

[10] Joseph W Goodman. *Introduction to Fourier optics*. Macmillan Learning, 4 edition, 2017. 3

[11] Xiuye Gu, Weixin Luo, Michael S Ryoo, and Yong Jae Lee. Password-conditioned anonymization and deanonymization with face identity transformers. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 6

[12] 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, 3, 7

[13] Carlos Hinojosa, Miguel Marquez, Henry Arguello, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Privhar: Recognizing human actions from privacy-preserving lens. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 3, 7

[14] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*. Springer, 2019. 1, 2, 6

[15] Baowei Jiang, Bing Bai, Haozhe Lin, Yu Wang, Yuchen Guo, and Lu Fang. Dartblur: Privacy preservation with detection artifact suppression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[16] 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. 5

[17] 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. 5

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[19] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 7

[20] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6

[21] Sudhakar Kumawat and Hajime Nagahara. Privacy-preserving action recognition via motion difference quantization. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[22] Vasudevan Lakshminarayanan and Andre Fleck. Zernike polynomials: a guide. *Journal of Modern Optics*, 2011. 3

[23] Dongze Li, Wei Wang, Kang Zhao, Jing Dong, and Tieniu Tan. Riddle: Reversible and diversified de-identification with latent encryptor. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 5, 6, 8

[24] Chi Liu, Tianqing Zhu, Jun Zhang, and Wanlei Zhou. Privacy intelligence: A survey on image privacy in online social networks. *ACM Computing Surveys*, 2022. 1, 2

[25] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 4

[26] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems (NeurIPS)*, 2018. 4

[27] Yuchen Luo, Junwei Zhu, Keke He, Wenqing Chu, Ying Tai, Chengjie Wang, and Junchi Yan. Styleface: Towards identity-disentangled face generation on megapixels. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[28] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[29] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixe. Ciagan: Conditional identity anonymization generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[30] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adver-

sarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[31] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *the 35th International Conference on Machine Learning*. PMLR, 2018. 4

[32] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference. Proceedings, Part VIII 14*. Springer, 2016. 4

[33] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[34] Francesco Pittaluga and Sanjeev J Koppal. Privacy preserving optics for miniature vision sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[35] Francesco Pittaluga and Sanjeev Jagannatha Koppal. Precapture privacy for small vision sensors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016. 2

[36] Siddharth Ravi, Pau Climent-Pérez, and Francisco Florez-Revuelta. A review on visual privacy preservation techniques for active and assisted living. *Multimedia Tools and Applications*, 2023. 1, 2

[37] Michael Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *AAAI conference on artificial intelligence*, 2017. 2

[38] Michael Ryoo, Kiyoon Kim, and Hyun Yang. Extreme low resolution activity recognition with multi-siamese embedding learning. In *AAAI conference on artificial intelligence*, 2018. 1, 2

[39] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5

[40] 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, 3

[41] Vinkle Srivastav, Afshin Gangi, and Nicolas Padoy. Human pose estimation on privacy-preserving low-resolution depth images. In *International conference on medical image computing and computer-assisted intervention (MICCAI)*. Springer, 2019. 1

[42] 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 *European Conference on Computer Vision (ECCV)*. Springer, 2022. 1, 3

[43] Razvan Viorescu et al. 2018 reform of eu data protection rules. *European Journal of Law and Public Administration*, 2017. 1

[44] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[45] Hui-Po Wang, Tribhuvanesh Orekondy, and Mario Fritz. Infoscrub: Towards attribute privacy by targeted obfuscation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[46] Ping Wang, Lishun Wang, and Xin Yuan. Deep optics for video snapshot compressive imaging. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[47] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 4, 5

[48] Yunqian Wen, Bo Liu, Jingyi Cao, Rong Xie, and Li Song. Divide and conquer: a two-step method for high quality face de-identification with model explainability. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 5, 6

[49] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 5

[50] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 2016. 6

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[52] Ruoyu Zhao, Yushu Zhang, Tao Wang, Wenying Wen, Yong Xiang, and Xiaochun Cao. Visual content privacy protection: A survey. *arXiv preprint arXiv:2303.16552*, 2023. 1, 2