Supplementary Material
PrivHAR: Recognizing Human Actions From Privacy-preserving Lens

Carlos Hinojosa[1,2,*], Miguel Marquez[1], Henry Arguello[1], Ehsan Adeli[2], Li Fei-Fei[2], and Juan Carlos Niebles[2]

[1] Universidad Industrial de Santander, Colombia
[2] Stanford University, USA
https://carloshinojosa.me/project/privhar/

In this supplementary document, we include:

1. Light propagation and optics modeling.
2. PSF frequency analysis.
3. Face recognition results.
4. Precision-recall and ROC curves.
5. Deconvolution attacks details.
6. Hardware experiments.
7. Creators and license of the assets used in our paper.
8. Potential negative impact of our work.
9. Personal data/Human subjects discussion.

Please see the supplementary **video** for more qualitative results and failure cases of our proposed PrivHAR network (with Rubiksnet backbone).

1 Light propagation and image formation model

We adopt the same image formation model as in previous works [2,18,10]. Specifically, we model the light transport in the camera using a differentiable Fourier optics model [7].

Figure 1 depicts our optical system, which consists of a camera with two thin convex lenses and a phase mask (ϕ) between them. Assuming that the thin lens has a focal length z 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: $1/z = 1/d_1 + 1/d_2$. Therefore, an object at a distance d_1 in front of the lens appears in focus at a distance d_2 behind the lens. Assuming that the scene is at optical infinity, 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:

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

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

[*] carlos.hinojosa@saber.uis.edu.co

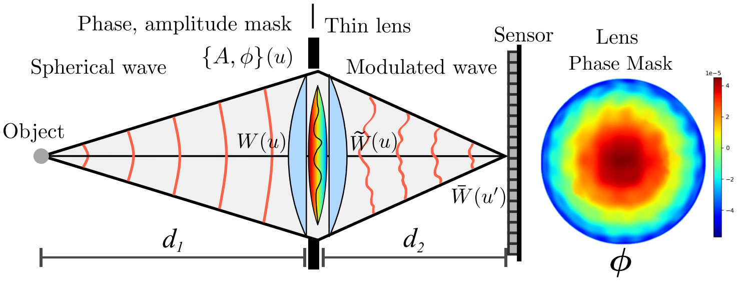


Fig. 1. Schematic diagram of the light propagation from the object at to the sensor with the focal length d_2. The phase of the spherical light wave coming from a scene point is modulated by our designed phase mask and captured by the camera's sensor. We take the magnitude-square of the light intensity measured by the sensor to find the values of the PSF **H**.

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 [7]

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

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:

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

Finally, the field propagates a distance d_2 to the sensor with the transfer function [7]:

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

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 **H** at each position (u, v) as:

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

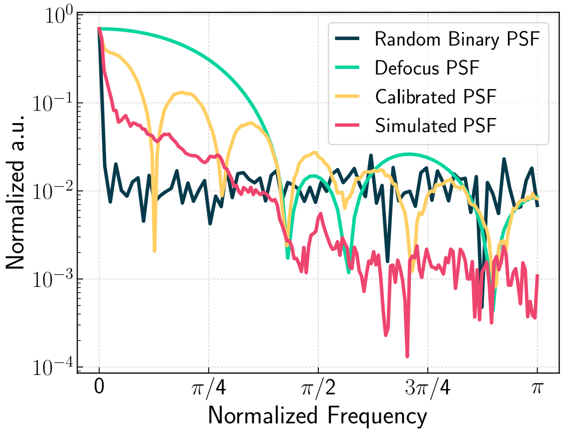


Fig. 2. Modulation Transfer Function (MTF) [1] of the point-spread functions (PSFs). The MTF is computed as the radially averaged magnitude spectrum of the PSF. The PSFs compared are: Defocus PSF [16], random binary, our simulated PSF, and our calibrated PSF of our proof-of-concept system. The magnitude spectrum of the proposed PSF decreases significantly for the entire frequency range indicating low invertibility characteristics, especially in the high-frequency range.

2 PSF frequency analysis.

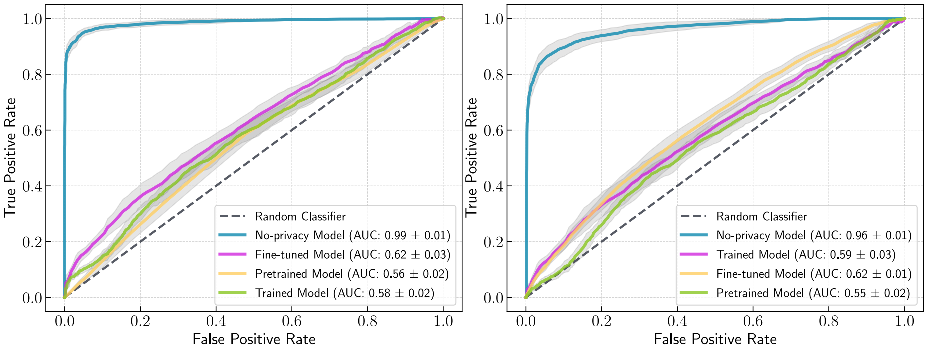
We additionally validate our proposed PSF using the modulation transfer function (MTF) metric [1]. The MTF is computed as the radially averaged magnitude spectrum of the PSF. As observed in Fig. 2, the magnitude spectrum of the proposed PSF decreases significantly for the entire frequency range indicating low invertibility characteristics, especially in the high-frequency range. This explains why our private images are more robust against face detection (high-frequency textures) than skin color detection attacks. Also, observe that the simulated PSF has lower invertibility than calibrated PSF. This is expected due to our real system's limitations.

3 Face recognition results.

In this work, we also measure privacy using face recognition. We use a Tensorflow implementation[‡] of the Additive Angular Margin Loss for Deep Face Recognition (ArcFace) network[3]. ArcFace is a recently published, efficient, and highly effective face recognition network that incorporates margins in its loss function to obtain highly discriminative features for face recognition.

We use the ArcFace network to test the face recognition performance on images acquired with our optimized lens. We experimented on three datasets: LFW [11], AgeDB-30 [14], and CFP-FP datasets[17]. We generate ROC curves

[‡] https://github.com/peteryuX/arcface-tf2



(a) AgeDB-30 Dataset (b) CFP-FP Dataset

Fig. 3. Face recognition performance on images from (a) AgeDB-30 and (b) CFP-FP datasets acquired with our optimized lens.

using three testing approaches for each dataset and compare them with the original ArcFace model tested on the original "non-private" images. We refer to the first approach as the "Pretrained model", which uses the pretrained ArcFace model to test the "private" version of each dataset. The second approach consists of training the ArcFace model from scratch using the private version of the MS-Celeb-1M dataset; we refer to such an approach as the "Trained model". Finally, in the "Finetuned model" approach, we first load the pretrained weights of the ArcFace model on original "non-private" images; then, we performed fine-tuning on the network with the private version of the MS-Celeb-1M dataset. We presented the results on the LFW dataset in Fig. 3 (b) of the main manuscript, and the results on the AgeDB-30 and CFP-FP dataset are shown in Fig. 3. Similar to the main manuscript results, the ArcFace model performs poorly on the "private" images generated by our optimized lens.

4 Precision-recall and ROC curves

To analyze the performance of the adversarial networks, we plot the receiver operating characteristic (ROC) and Precision-Recall (PR) curves. In Fig. 4, we show the ROC and PR curves of the adversarial network, which achieves the best performance on the privacy-preserving images/videos acquired with PrivHAR when using Rubiksnet as the backbone. For each curve, we show the area under the curve (AUC), the random classifier (null hypothesis), and perfect classifier performance for reference. We depict the random and perfect classifier with dashed and solid gray lines, respectively. As observed, the AUC values of the PR curves are very close to those obtained by a random classifier. Therefore, based on Fisher's exact test [20], the best adversarial network on our privacy-protected images is not significantly different from the random classifier (p-value < 0.01). In Fig. 5, we also show the corresponding PR curves obtained when using C3D as the backbone for PrivHAR. We observed a similar behavior, i.e.,

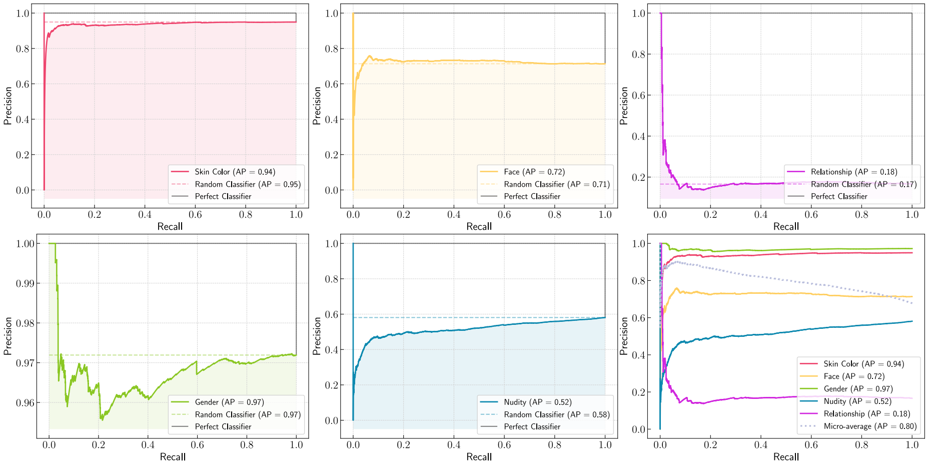


Fig. 4. Precision-recall curves when using Rubiksnet as backbone for PrivHAR.

the performance of A_A is close to the random classifier. In addition, we also show the receiver operating characteristic curves (ROC) in Fig. 6 when using (a) Rubiksnet and (b) C3D as the backbone in our PrivHAR network. However, note that the PA-HMDB51 dataset is imbalanced in the privacy attributes, and ROC Curves can be optimistic on severely imbalanced classification problems with few samples of the minority class [6,8]. Therefore, we made our principal analysis of the performance of A_A using PR curves in the main paper.

5 Deconvolution attacks details

In this work, we investigate the robustness of our proposed phase mask to deconvolution attacks. In general, there are 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 a more realistic 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. 5 of the main paper.

To test the robustness of our designed lens to blind deconvolution attacks, we trained a deconvolution network (DeblurGAN [13]) with 37608 sharp and blur (ours) images/frames from the HMDB51 dataset acquired with PrivHAR using C3D and Rubiksnet backbones. We use the same default parameters for DeblurGAN and train the network during 300 epochs. As observed from the results in the paper, reconstruction is challenging. The network can reconstruct some objects; however, the face details are missed, and the network cannot recover people's identities.

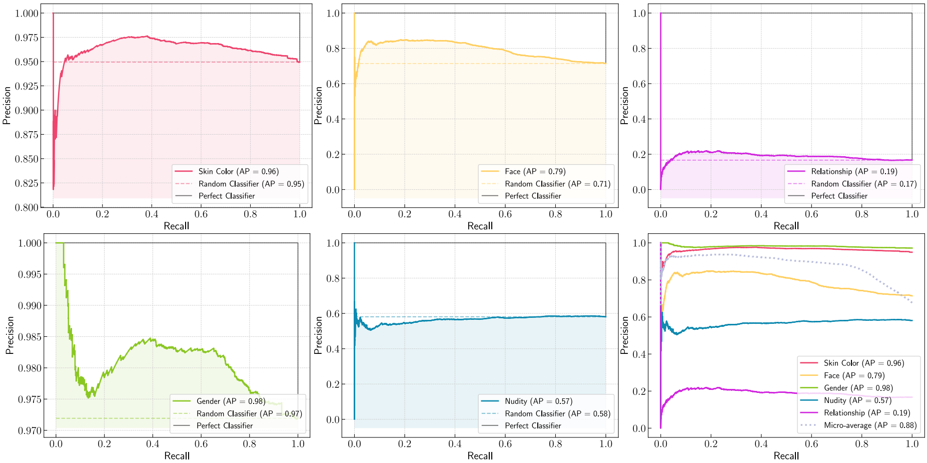
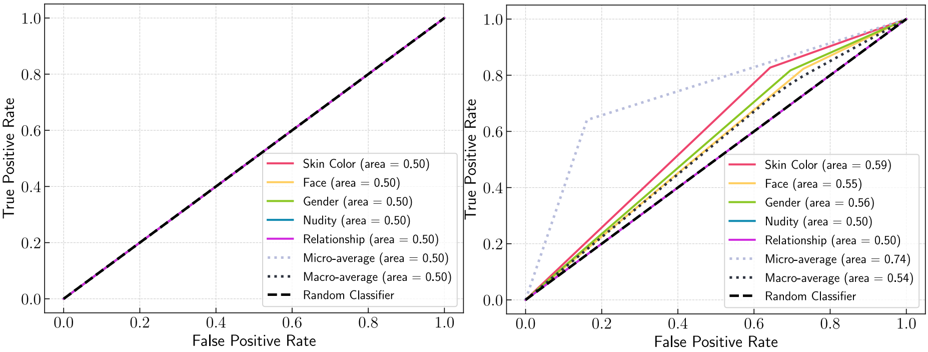


Fig. 5. Precision-recall curves when using C3D as backbone for PrivHAR.



(a) Rubiksnet

(b) C3D

Fig. 6. ROC curves when using (a) Rubiksnet, and (b) C3D as backbone for PrivHAR.

On the other hand, we also use a non-blind deconvolution approach (Wiener deconvolution [4]) to try to recover the underlying scene. This approach also does not work well for our proposed lens design, especially for a lens designed with Rubiksnet as HAR backbone, as it has significantly more aberrations, making it more robust.

6 Hardware experiments

Optical Architecture. Our proof-of-concept system uses an objective CANON lens (CANON, EF-S 18-55 mm f/4-5.6 IS STM) to image the scene to an intermediate plane. Then, the intermediate image plane is relayed onto two CANON camera sensors (EOS M50, 24.1 MP APS-C) by a double 4f system consisting of

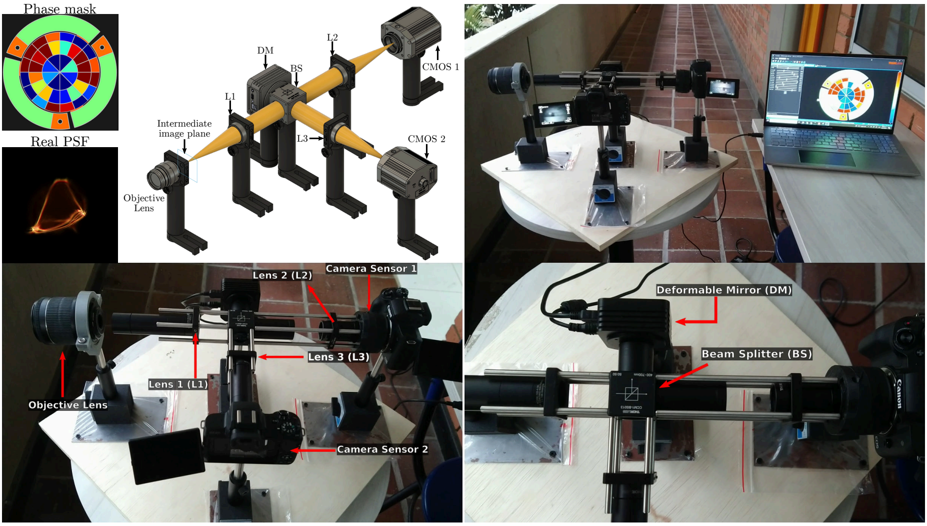


Fig. 7. Proof-of-concept optical system. We build our optical architecture with two CANON cameras, one objective CANON lens, three lenses, one beam splitter, and one deformable mirror. For more details please see the supplementary video.

Lens 1 (L1), Lens 2 (L2), Lens 3 (L3) (Thorlabs, AC254-100-A), a beam splitter (BS) (Thorlabs, BS013 - 50:50), and a deformable mirror (DM) (Thorlabs, DMP40-P01). We place the BS after Lens 1 to split and relay the incoming wavefront to camera sensors 1 and 2. In this way, camera sensor 1 acquires the No-privacy videos. We place the DM on the Fourier plane of Lens 1 and 3 to introduce the spatial deformation and reflect the phase-encoded wavefront to Lens 3. Finally, the phase-encoded scene is integrated by the camera sensor 2 and acquires privacy-protected video with a size of up to 6000 × 4000 pixels ($3.717\mu m$ $\times 3.717$ μm). To calibrate the point spread function (PSF) induced by the DM, we place a fiber optics (Ocean Insight, QR200-7-UV-BX) with a kernel core of 200 μm at $1m$ from the CANON lens. Figure 7 shows pictures of our proof-of-concept system with two cameras. For more details, please see the supplementary video.

Quantitative Results. After calibration, we captured 282 videos in total, 141 privacy-protected, and 141 no-privacy videos, with eight different persons (see section 9). We use a small set of captured privacy-protected videos with the respective action ground truth labels to fine-tune our PrivHAR network for 30 epochs. Specifically, we use 75 privacy-protected videos for fine-tuning and 66 videos for testing. We obtained an action recognition accuracy of $A_C = 83.32$ on the testing set. Due to pandemic restrictions, we cannot acquire a larger-scale video dataset in the lab. For now, our small-scale tests show results consistent with our extensive experiments. As future work, we plan to build a privacy-preserving video dataset using our proposed optical system, which allows us to acquire both RGB and privacy-protected videos.

Table 1. Assets descriptions used in our work.

Asset Name	Type	Reference	Implementation	License
Rubiksnet	HAR Backbone	[5]	https://tinyurl.com/rubiksnet	MIT
C3D	HAR Backbone	[19]	https://tinyurl.com/c3dbackbone	MIT
PA-HMDB51	Dataset	[21]	https://tinyurl.com/pa-hmdb51	Licensed Material
HMDB51	Dataset	[12]	https://tinyurl.com/hmdb51	CC BY 4.0
VISPR		[15]	https://tinyurl.com/vispr-dataset	CC BY 4.0
DeblurGAN	Neural Network	[13]	https://tinyurl.com/deblurganv2	BSD
ResNet-50	Neural Network	[9]	https://tinyurl.com/resnet-pytorch	BSD 3-Clause
ArcFace	Neural Network	[9]	https://github.com/peteryuX/arcface-tf2	MIT

Qualitative Results. Please see the supplementary video to see qualitative results on acquired privacy-preserving videos with our proof-of-concept optical system.

7 Creators and license of the assets used in our paper

We appropriately cited all the assets (datasets and codes) in our main paper and supplementary document. In Table 1, we summarize the type, reference, used implementation (or website), and license of the main assets used in this work.

8 Potential negative impact of our work

In this work, we aim at addressing one social concern in the vision community: the development of privacy-preserving vision systems. We explore the design of a lens for protecting privacy while performing human action recognition. Although we did not identify a direct potential negative impact of our work, we think that one attacker could adopt a similar approach and develop or simulate a camera to acquire adversarial examples and attack other HAR networks to decrease their performance.

9 Personal data/Human subjects discussion

In this work, we acquired videos from people doing actions in our Lab. In total, eight subjects with ages between 20 and 26 years old collaborated in the acquisition of the videos with our prototype camera described in section 6. 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. Boominathan, V., Adams, J.K., Robinson, J.T., Veeraraghavan, A.: Phlatcam: Designed phase-mask based thin lensless camera. IEEE transactions on pattern analysis and machine intelligence **42**(7), 1618–1629 (2020) 3
2. Chang, J., Sitzmann, V., Dun, X., Heidrich, W., Wetzstein, G.: Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. Scientific reports **8**(1), 1–10 (2018) 1
3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019) 3
4. Dong, J., Roth, S., Schiele, B.: Deep wiener deconvolution: Wiener meets deep learning for image deblurring. Advances in Neural Information Processing Systems **33**, 1048–1059 (2020) 6
5. Fan, L., Buch, S., Wang, G., Cao, R., Zhu, Y., Niebles, J.C., Fei-Fei, L.: Rubiksnet: Learnable 3d-shift for efficient video action recognition. In: European Conference on Computer Vision. pp. 505–521. Springer (2020) 8
6. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from imbalanced data sets, vol. 10. Springer (2018) 5
7. Goodman, J.W.: Introduction to Fourier optics. Macmillan Learning, 4 edition (2017) 1, 2
8. He, H., Ma, Y.: Imbalanced learning: foundations, algorithms, and applications. Wiley-IEEE Press (2013) 5
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 8
10. Hinojosa, C., Niebles, J.C., Arguello, H.: Learning privacy-preserving optics for human pose estimation. In: ICCV. pp. 2573–2582 (October 2021) 1
11. Huang, G.B., Mattar, M., Lee, H., Learned-Miller, E.: Learning to align from scratch. In: NIPS (2012) 3
12. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011) 8
13. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8878–8887 (2019) 5, 8
14. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (2017) 3
15. Orekondy, T., Schiele, B., Fritz, M.: Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In: Proceedings of the IEEE international conference on computer vision. pp. 3686–3695 (2017) 8
16. Pittaluga, F., Koppal, S.J.: Privacy preserving optics for miniature vision sensors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 314–324 (2015) 3
17. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2016) 3

18. Sitzmann, V., Diamond, S., Peng, Y., Dun, X., Boyd, S., Heidrich, W., Heide, F., Wetzstein, G.: End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. ACM TOG (2018) 1
19. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015) 8
20. Upton, G.J.: Fisher's exact test. Journal of the Royal Statistical Society: Series A (Statistics in Society) **155**(3), 395–402 (1992) 4
21. Wu, Z., Wang, H., Wang, Z., Jin, H., Wang, Z.: Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 8