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], Juan Carlos Niebles[2]

[1]Universidad Industrial de Santander, [2]Stanford University

carlos.hinojosa@saber.uis.edu.co

High Dimensional Signal Processing Research Group

Project Page

ECCV TEL AVIV 2022

Motivation

Cameras are everywhere! How to develop privacy-preserving vision systems?

We want to have smart systems (e.g., in our homes, hospitals, etc.) to recognize relevant events and assist us in our activities. However, we also want to ensure **they protect our privacy.**

Prior work on Privacy-preserving vision

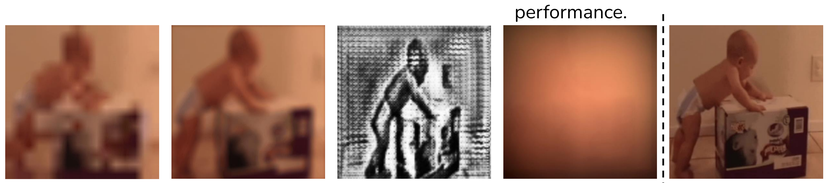
Low-resolution
- Lose information.
- Action recognition decreases.

Defocus
- Susceptible to reverse engineering attacks.

PDAR
- **Software-level** method.
- Use traditional cameras.

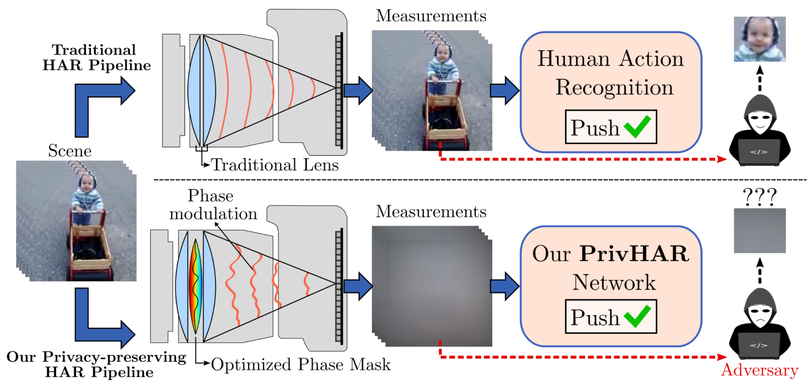
Ours
- **Hardware-level** method.
- We learn the distortion to preserve HAR performance.

No-Privacy



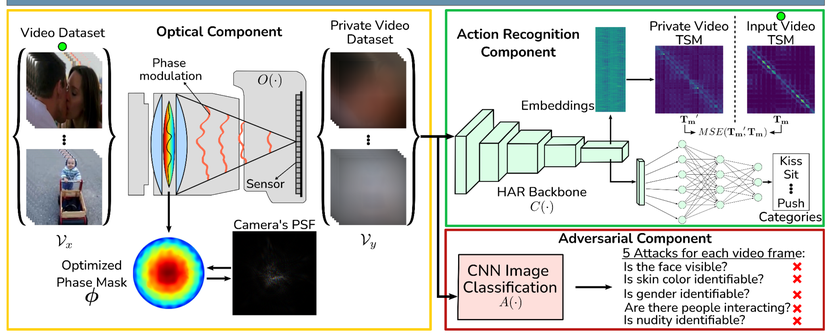
Our key idea: instead of fixed/manually defined optics, we'll design optical distortion in a way that doesn't degrade HAR performance.

Traditional HAR Pipeline vs PrivHAR (Ours)



In the traditional HAR pipeline, commercial cameras acquire visual details from the scene leading to privacy issues. We introduce **PrivHAR**, an adversarial optimization framework that learns a lens' phase mask to encode human action features and perform HAR while obscuring privacy-related attributes.

Model and Approach

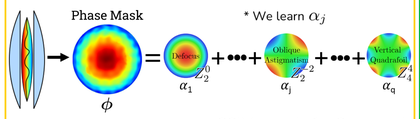


Our camera comprises two thin convex lenses and a phase mask between them. We propose a framework to accomplish three goals:

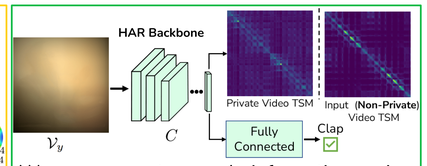
1. To learn to add aberrations to the lens surface such that the acquired videos are distorted to obscure private attributes but still preserve important features.
2. To learn the parameters of an action recognition network to perform HAR on the private videos with high accuracy.
3. To obtain private videos that are robust to adversarial attacks.

Adversarial Optimization

We parameterize the surface profile of the lens with Zernike polynomials, where each one describes a wavefront aberration.

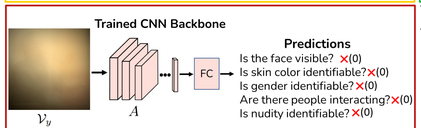


$L(O) \triangleq SSIM(\mathbf{X}, \mathbf{Y})$ SSIM stands for Structural Similarity Index Measure.



We preserve temporal information using temporal similarity matrices (**TSMs**).

$L(C) \triangleq \mathcal{H}(\mathcal{S}_C, C(\mathcal{V}_y)) + MSE(\mathbf{T_m}, \mathbf{T_m}')$



We define the adversarial attack as a multilabel binary classification problem: Attribute **identifiable (1) / unidentifiable (0)**.

$L(A) \triangleq \mathcal{H}(\mathcal{S}_A, A(\mathcal{V}_y))$

Algorithm 1: Our Adversarial Training Algorithm.

Input : Video Dataset $\mathcal{V}_x = \{\mathbf{V_x}^e\}_{e=1}^E$.
Hyperparameters $\beta_o, \beta_c, \beta_a, \gamma_1, \gamma_2$
Output: $\theta_o, \theta_c, \theta_a$
Function $Train(\mathcal{V}_x, \beta_o, \beta_c, \beta_a, \gamma_1, \gamma_2)$

```
for every epoch do
    for every batch of videos V_x^B do
        V_y^B = O(V_x^B) ▷ Acquire private videos
        θ_o ← θ_o − β_o Δ_θo (L(O) + γ1 L(C) − γ2 L(A))
        θ_c ← θ_c − β_c Δ_θc (L(C))
        θ_a ← θ_a − β_a Δ_θa (L(A))
return X_e
```

Datasets and Metrics

We performed cross-dataset training using three datasets: the **HMDB51**, **VISPR**, and the **PA-HMDB51**. We use PA-HMDB51 for **testing** our action recognition and adversarial components. We used five privacy attributes: **skin color, face, gender, nudity, and relationship**. We used the following metrics in our framework:

HAR ↑	Image Quality ↓	Adversarial Component ↓	Face Recognition ↓
We report the standard average classification accuracy, denoted by **AC**	We use the structural similarity index measure (**SSIM**) to measure image degradation. We expect low SSIM values.	We adopt the Class based Mean Average Precision **(C-MAP or AA)** to evaluate the adversarial models.	We measure Face recognition in terms of the area under the curve (**AUC**) of the ROC.

Quantitative Results

Ablation Study

C3D Backbone

Experiment	SSIM↓	A_C ↑	A_A ↓	P ↑
No-Adversarial	0.603	51.1	69.1	38.6
No-TSM	0.612	59.9	69.7	40.2
Zernike-50	0.643	58.3	70.5	39.2
Zernike-100	0.629	58.8	69.3	40.4
Zernike-200	0.612	**63.3**	**68.9**	**41.52**

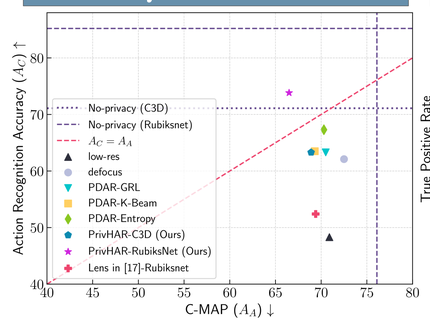
RubiksNet Backbone

Experiment	SSIM↓	A_C ↑	A_A ↓	P ↑
No-Adversarial	0.592	57.6	68.2	40.9
No-TSM	0.599	72.3	67.6	44.6
Zernike-50	0.618	70.2	69.2	42.8
Zernike-100	0.601	71.9	68.4	43.9
Zernike-200	0.588	**73.8**	**66.5**	**46.1**

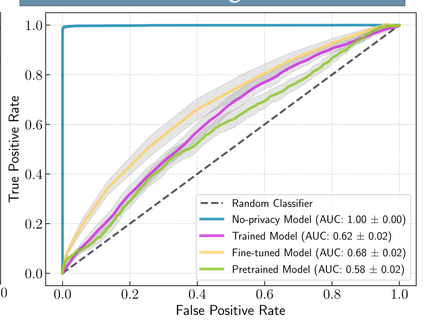
Comparisons

Methods	SSIM↓	A_C ↑	A_A ↓	P ↑
No-privacy (C3D)	1.0	71.1	76.1	35.8
No-privacy (RubiksNet)	1.0	85.2	76.1	37.3
Low-resolution [41]	0.686	48.3	70.9	36.3
Lens in [17]-RubiksNet	0.608	52.4	69.4	38.6
Defocus [37]	0.688	62.1	72.5	38.1
PDAR-GRL [56]	-	63.3	70.5	40.2
PDAR-K-Beam [56]	-	63.5	69.3	41.4
PDAR-Entropy [56]	-	67.3	70.3	41.2
PrivHAR-C3D	0.612	63.3	68.9	41.7
PrivHAR-RubiksNet	0.588	**73.8**	**66.5**	**46.1**

Privacy / HAR Trade-off



Face Recognition



Qualitative Results