Privacy-Preserving Deep Action Recognition: An Adversarial Learning Framework and A New Dataset

Zhenyu Wu*, Haotao Wang*, Zhaowen Wang, Hailin Jin, and Zhangyang Wang

Abstract—We investigate privacy-preserving, video-based action recognition in deep learning, a problem with growing importance in smart camera applications. A novel adversarial training framework is formulated to learn an anonymization transform for input videos such that the trade-off between target utility task performance and the associated privacy budgets is explicitly optimized on the anonymized videos. Notably, the privacy budget, often defined and measured in task-driven contexts, cannot be reliably indicated using any single model performance because strong protection of privacy should sustain against any malicious model that tries to steal private information. To tackle this problem, we propose two new optimization strategies of model restarting and model ensemble to achieve stronger universal privacy protection against any attacker models. Extensive experiments have been carried out and analyzed.

On the other hand, given few public datasets available with both utility and privacy labels, the data-driven (supervised) learning cannot exert its full power on this task. We first discuss an innovative heuristic of cross-dataset training and evaluation, enabling the use of multiple single-task datasets (one with target task labels and the other with privacy labels) in our problem. To further address this dataset challenge, we have constructed a new dataset, termed PA-HMDB51, with both target task labels (action) and selected privacy attributes (gender, age, race, nudity, and relationship) annotated on a per-frame basis. This first-of-its-kind video dataset and evaluation protocol can greatly facilitate visual privacy research and open up other opportunities. Our codes, models, and the PA-HMDB51 dataset are available at: *https://github.com/VITA-Group/PA-HMDB51*.

Index Terms—Visual privacy, action recognition, privacy-preserving learning, adversarial learning.

1 Introduction

Smart surveillance or smart home cameras, such as Amazon Echo and Nest Cam, are now found in millions of locations to link users to their homes or offices remotely. They provide the users with a monitoring service by notifying environment changes, a lifelogging service, and intelligent assistance. However, the benefits come at the heavy price of privacy intrusion from time to time. Due to the computationally demanding nature of visual recognition tasks, only some can run on the resource-limited local devices, which makes transmitting (part of) data to the cloud necessary. Growing concerns have been raised [1]–[4] towards personal data uploaded to the cloud, which could be potentially misused or stolen by malicious third-parties. This new privacy risk is fundamentally different from traditional concerns over unsecured transmission channels (*e.g.*, malicious third-party eavesdropping), and therefore requires new solutions to address it. Many laws and regulations in the United States and the European Union [5]–[8] also bring up guidelines for handling *Personally Identifiable Information*.

We ask if it is possible to alleviate privacy concerns without compromising user convenience. At first glance, the question itself is posed as a dilemma: we would like a camera system to recognize important events and assist daily human life by understanding its videos while preventing it from obtaining sensitive visual information (such as faces, gender, skin color, etc.) that can intrude individual privacy. Thus, it becomes a new and appealing problem to find an appropriate transform to obfuscate the captured raw visual data at the local end, so that the transformed data will only enable specific target utility tasks while obstructing undesired privacy-related tasks. Recently, some new video acquisition approaches [9]–[11] were proposed to intentionally capture or process videos in extremely low-resolution to create privacy-preserving "anonymized" videos and showed promising empirical results.

This paper takes one of the first steps towards addressing this challenge of privacy-preserving, video-based action recognition, via the following contributions:

- **A General Adversarial Training and Evaluation Framework.** We address the privacy-preserving action recognition problem with a novel adversarial training framework. The framework explicitly optimizes the trade-off between target utility task performance and the associated privacy budgets by learning to anonymize the original videos. To reduce the training instability (as discussed in Section 3.1), we design and compare three different optimization strategies. We empirically find one strategy generally outperforms the others under our framework

- *Zhenyu Wu is with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX, 77840. E-mail: wuzhenyu_sjtu@tamu.edu*
- *Haotao Wang and Zhangyang Wang are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, TX, 78712. E-mail: {htwang, atlaswang}@utexas.edu*
- *Zhaowen Wang and Hailin Jin are with Adobe Research, San Jose, CA, 95110. E-mail: {zhawang, hljin}@adobe.com*
- *The first two authors Zhenyu Wu and Haotao Wang contributed equally to this work.*
- *Correspondence to Zhangyang Wang (atlaswang@utexas.edu).*

and provide intuition to its advantage.

- **Practical Approximations of "Universal" Privacy Protection.** The privacy budget in our framework cannot be defined *w.r.t.* one model that predicts privacy attributes. Instead, the ideal protection of privacy must be universal and model-agnostic, *i.e.*, preventing every possible attacker model from predicting private information. To resolve this so-called **"$\forall$ challenge"**, we propose two effective strategies, *i.e.*, *restarting* and *ensembling*, to enhance the generalization capability of the learned anonymization to defend against unseen models. We leave it as our future work to find better methods for this challenge.
- **A New Dataset with Action and Privacy Annotations.** When it comes to evaluating privacy protection on complicated privacy attributes, there is no off-the-shelf video dataset with both action (utility) and privacy attributes annotated, either for training or testing. Such a dataset challenge is circumvented in our previous work [12] by using the VISPR [13] dataset as an auxiliary dataset to provide privacy annotations for cross-dataset evaluation (details in Section 3.6). However, this protocol inevitably suffers from the domain gap between the two datasets: while the utility was evaluated on one dataset, the privacy was measured on a different dataset. The incoherence in utility and privacy evaluation datasets makes the obtained utility-privacy trade-off less convincing. To reduce this gap, in this paper, we construct the very first testing benchmark dataset, dubbed **P**rivacy **A**nnotated **HMDB51** (PA-HMDB51), to evaluate privacy protection and action recognition on the same videos simultaneously. The new dataset consists of 515 videos originally from HMDB51. For each video, privacy labels (five attributes: skin color, face, gender, nudity, and relationship) are provided on a per-frame basis. We benchmark our proposed framework on the new dataset and validate its effectiveness.

The paper is built upon our prior work [12] with multiple improvements: (1) A detailed discussion and comparison on three optimization strategies for the proposed framework; (2) A much expanded experimental and analysis section; and (3) most importantly, the construction of the new PA-HMDB51 dataset, and the associated benchmark results.

2 Related Work

2.1 Privacy Protection in Computer Vision

With pervasive cameras for surveillance or smart home devices, privacy-preserving action recognition has drawn increasing interests from both industry and academia.

Transmitting Feature Descriptors A seemingly reasonable and computationally cheaper option is to extract feature descriptors from raw images and transmit those features only. Unfortunately, previous studies [14]–[18] revealed that considerable details of original images could still be recovered from standard HOG, SIFT, LBP, 3D point clouds, Bag-of-Visual-Words or neural network activations (even if they look visually distinctive from natural images).

Homomorphic Cryptographic Solutions Most classical cryptographic solutions secure communication against unauthorized access from attackers. However, they are not immediately applicable to preventing authorized agents (such as the back-end analytics) from the unauthorized abuse of information, causing privacy breach concerns. A few encryption-based solutions, such as Homomorphic Encryption (HE) [19], [20], were developed to locally encrypt visual information. The server can only get access to the encrypted data and conduct a utility task on it. However, many encryption-based solutions will incur high computational costs at local platforms. It is also challenging to generalize the cryptosystems to more complicated classifiers. Chattopadhyay *et al*. [21] combined the detection of regions of interest and the real encryption techniques to improve privacy while allowing general surveillance to continue.

Anonymization by Empirical Obfuscations An alternative approach towards a privacy-preserving vision system is based on the concept of anonymized videos. Such videos are intentionally captured or processed by empirical obfuscations to be in special low-quality conditions, which only allow for recognizing some target events or activities while avoiding the unwanted leak of the identity information for the human subjects in the video.

Ryoo *et al*. [9] showed that even at the extremely low resolutions, reliable action recognition could be achieved by learning appropriate downsampling transforms, with neither unrealistic activity-location assumptions nor extra specific hardware resources. The authors empirically verified that conventional face recognition easily failed on the generated low-resolution videos. Butler *et al*. [10] used image operations like blurring and superpixel clustering to get anonymized videos, while Dai *et al*. [11] used extremely low resolution (*e.g.*, 16×12) camera hardware to get anonymized videos. Winkler *et al*. [22] used cartoon-like effects with a customized version of mean shift filtering. Wang *et al*. [23] proposed a lens-free coded aperture (CA) camera system, producing visually unrecognizable and unrestorable image encodings. Pittaluga & Koppal [24], [25] proposed to use privacy-preserving optics to filter sensitive information from the incident light-field before sensor measurements are made, by k-anonymity and defocus blur. Earlier work of Jia *et al*. [26] explored privacy-preserving tracking and coarse pose estimation using a network of ceiling-mounted time-of-flight low-resolution sensors. Tao *et al*. [27] adopted a network of ceiling-mounted binary passive infrared sensors. However, both works [26], [27] handled only a limited set of activities performed at specific constrained areas in the room.

The usage of low-quality anonymized videos by obfuscations was computationally cheap and compatible with sensor and bandwidth constraints. However, the proposed obfuscations were not learned towards protecting any visual privacy, thus having limited effects. In other words, privacy protection came as a "side product" of obfuscation, and was not a result of any optimization, making the privacy protection ability very limited. What is more, the privacy-preserving effects were not carefully analyzed and evaluated by human study or deep learning-based privacy recognition approaches. Lastly, none of the aforementioned empirical obfuscations extended their efforts to study deep learning-based action recognition, making their task performance less competitive. Similarly, the recent progress of low-resolution object recognition [28]–[30] also put their privacy protection effects in jeopardy.

Learning-based Solutions Very recently, a few learning-

based approaches have been proposed to address privacy protection or fairness problems in vision-related tasks [12], [31]–[39]. Many of them exploited ideas from adversarial learning. They addressed this problem by learning data representations that simultaneously reduce the budget cost of privacy or fairness while maintaining the utility task performance.

Wu *et al*. [31] proposed an adversarial training framework dubbed Nuisance Disentangled Feature Transform (NDFT) to utilize the free meta-data (*i.e*., altitudes, weather conditions, and viewing angles) in conjunction with associated UAV images to learn domain-robust features for object detection in UAV images. Pittaluga *et al*. [33] preserved the utility by maintaining the variance of the encoding or favoring a second classifier for a different attribute in training. Bertran *et al*. [34] motivated the adversarial learning framework as a distribution matching problem and defined the objective and the constraints in mutual information. Roy & Boddeti [35] measured the uncertainty in the privacy-related attributes by the entropy of the discriminator's prediction. Oleszkiewicz *et al*. [40] proposed an empirical data-driven privacy metric based on mutual information to quantify the privatization effects on biometric images. Zhang *et al*. [36] presented an adversarial debiasing framework to mitigate the biases concerning demographic groups. Ren *et al*. [37] learned a face anonymizer in video frames while maintaining the action detection performance. Shetty *et al*. [38] presented an automatic object removal model that learns how to find and remove objects from general scene images via a generative adversarial network (GAN) framework.

2.2 Privacy Protection in Social Media/Photo Sharing

User privacy protection is also a topic of extensive interest in the social media field, especially for photo sharing. The most common means to protect user privacy in an uploaded photo is to add empirical obfuscations, such as blurring, mosaicing, or cropping out certain regions (usually faces) [41]. However, extensive research showed that such an empirical approach could be easily hacked [42], [43]. A recent work [44] described a game-theoretical system in which the photo owner and the recognition model strive for antagonistic goals of disabling recognition, and better obfuscation ways could be learned from their competition. However, their system was only designed to confuse one specific recognition model via finding its adversarial perturbations. Fooling only one recognition model can cause obvious overfitting as merely changing to another recognition model will likely put the learning efforts in vain: such perturbations cannot even protect privacy from human eyes. The problem setting in [44] thus differs from our target problem. Another notable difference is that we usually hope to generate minimum perceptual quality loss to photos after applying any privacy-preserving transform to them in social photo sharing. There is no such restriction in our scenario. We can apply a much more flexible and aggressive transformation to the image.

The visual privacy issues faced by blind people were revealed in [45] with the first dataset in this area. Concrete privacy attributes were defined in [13] with their correlation with image content. The authors categorized possible private information in images, and they ran a user study to understand privacy preferences. They then provided a sizable set of 22k images annotated with 68 privacy attributes, on which they trained privacy attributes predictors.

3 Method

3.1 Problem Definition

Objective Assume our training data $\mathcal{X}$ (raw visual data captured by camera) are associated with a target utility task $\mathcal{T}$ and a privacy budget $\mathcal{B}$. Since $\mathcal{T}$ is usually a supervised task, *e.g*., action recognition or visual tracking, a label set $\mathcal{Y}_T$ is provided on $\mathcal{X}$, and a standard cost function L_T (*e.g*., cross-entropy) is defined to evaluate the task performance on $\mathcal{T}$. Usually, there is a state-of-the-art deep neural network f_T, which takes $\mathcal{X}$ as input and predicts the target labels. On the other hand, we need to define a budget cost function J_B to evaluate its input data's privacy leakage: the smaller $J_B(\cdot)$ is, the less private information its input contains.

We seek an optimal anonymization function f_A^* to transform the original $\mathcal{X}$ to anonymized visual data $f_A^*(\mathcal{X})$, and an optimal target model f_T^* such that:

- f_A^* has filtered out the private information in $\mathcal{X}$, *i.e*.,

$$J_B(f_A^*(\mathcal{X})) \ll J_B(\mathcal{X})$$

- the performance of f_T is minimally affected when using the anonymized visual data $f_A^*(\mathcal{X})$ compared to when using the original data $\mathcal{X}$, *i.e*.

$$L_T(f_T^*(f_A^*(\mathcal{X})), \mathcal{Y}_T) \approx \min_{f_T} L_T(f_T(\mathcal{X}), \mathcal{Y}_T)$$

To achieve these two goals, we mathematically formulate the problem as solving the following optimization problem:

$$f_A^*, f_T^* = \underset{f_A, f_T}{\operatorname{argmin}} [L_T(f_T(f_A(\mathcal{X})), \mathcal{Y}_T) + \gamma J_B(f_A(\mathcal{X}))] \quad (1)$$

Definition of J_B and L_T The definition of the privacy budget cost J_B is not straightforward. Practically, it needs to be placed in concrete application contexts, often in a task-driven way. For example, in smart workplaces or smart homes with video surveillance, one might often want to avoid disclosure of the face or identity of persons. Therefore, to reduce J_B could be interpreted as to suppress the success rate of identity recognition or verification. Other privacy-related attributes, such as race, gender, or age, can be similarly defined too. We denote the privacy-related annotations (such as identity label) as $\mathcal{Y}_B$, and rewrite $J_B(f_A(\mathcal{X}))$ as $J_B(f_B(f_A(\mathcal{X})), \mathcal{Y}_B)$, where f_B denotes the privacy budget model which takes (anonymized or original) visual data as input and predicts the corresponding private information. Different from L_T, minimizing J_B will encourage $f_B(f_A(\mathcal{X}))$ to diverge from $\mathcal{Y}_B$. Without loss of generality, we assume both f_T and f_B to be classification models and output class labels. Under this assumption, we could choose L_T as the cross-entropy function and J_B as the negative cross-entropy function:

$$L_T \triangleq H(\mathcal{Y}_T, f_T(f_A(\mathcal{X})))$$
$$J_B \triangleq -H(\mathcal{Y}_B, f_B(f_A(\mathcal{X})))$$

where $H(\cdot,\cdot)$ is the cross-entropy function. For convenience, we also define another variable L_B as $-J_B$:

$$L_B \triangleq -J_B = H(\mathcal{Y}_B, f_B(f_A(\mathcal{X})))$$

Two Challenges Such a supervised, task-driven definition of J_B poses at least two challenges: (1) *Dataset challenge:* The privacy budget-related annotations, denoted as $\mathcal{Y}_B$, often have less availability than target utility task labels. Specifically, it is often challenging to have both $\mathcal{Y}_T$ and $\mathcal{Y}_B$ available on the same $\mathcal{X}$; (2) $\forall$ *challenge:* Considering the nature of privacy protection, it is not sufficient to merely suppress the success rate of one f_B model. Instead, we define a privacy prediction function family

$$\mathcal{P} : f_A(\mathcal{X}) \mapsto \mathcal{Y}_B,$$

so that the ideal privacy protection by f_A should be reflected as suppressing every possible model f_B from $\mathcal{P}$. That differs from the common supervised training goal, where only one model needs to be found to fulfill the target utility task successfully.

We address the *dataset challenge* by two ways: (1) cross dataset training and evaluation (Section 3.4); and more importantly (2) building a new dataset annotated with both utility and privacy labels (Section 5). We defer their discussion to respective experimental paragraphs.

Handling the $\forall$ *challenge* is more challenging. Firstly, we re-write the general form in Eq. (1) with the task-driven definition of J_B as follows:

$$f_A^*, f_T^* = \text{argmin}_{(f_A,f_T)}[L_T(f_T(f_A(\mathcal{X})), \mathcal{Y}_T) + \gamma \sup_{f_B \in \mathcal{P}} J_B(f_B(f_A(\mathcal{X})), \mathcal{Y}_B)]. \quad (2)$$

The $\forall$ *challenge* is the infeasibility to directly solve Eq. (2), due to the infinite search space of f_B in $\mathcal{P}$. Secondly, we propose to solve the following approximate problem by setting f_B as a neural network with a fixed structure:

$$f_A^*, f_T^* = \text{argmin}_{(f_A,f_T)}[L_T(f_T(f_A(\mathcal{X})), \mathcal{Y}_T) + \gamma \max_{f_B} J_B(f_B(f_A(\mathcal{X})), \mathcal{Y}_B)]. \quad (3)$$

Lastly, we propose "model ensemble" and "model restarting" (Section 3.5) to handle the $\forall$ *challenge* better and boost the experimental results further.

Considering the $\forall$ *challenge*, the evaluation protocol for privacy-preserving action recognition is more intricate than traditional action recognition task. We propose a two-step protocol (as described in Section 3.6) to evaluate f_A^* and f_T^* on the trade-off they have achieved between target task utility and privacy protection budget.

Solving the Minimax Solving Eq. (3) is still challenging because the minimax problem is hard by its nature. Traditional minimax optimization algorithms based on alternating gradient descent can only find minimax points for convex-concave problems, and they achieve sub-optimal solutions on deep neural networks since they are neither convex nor concave. Some very recent minimax algorithms, such as K-Beam [46], have been shown to be promising in none convex-concave and deep neural network applications. However, these methods rely on heavy parameter tuning and are effective only in limited situations. Besides, our optimization goal in Eq. (3) is even harder than common minimax objectives like those in GANs, which are often interpreted as a two-party competition game. In contrast, our Eq. (3) is more "hybrid" and can be interpreted as a more complicated three-party competition, where (adopting machine learning security terms) f_A is an obfuscator, f_T a utilizer collaborating with the obfuscator, and f_B an attacker trying to breach the obfuscator. Therefore, we see no obvious best choice from the off-the-shelf minimax algorithms to ahieve our objective.

We are thus motivated to try different state-of-the-art minimax optimization algorithms on our framework. We tested two state-of-the-art minimax optimization algorithms, namely GRL [47] and K-Beam [46], on our framework and proposed an innovative entropy maximization method to solve Eq. (3). We empirically show our entropy maximization algorithm outperforms both state-of-the-art minimax optimization algorithms and discuss its advantages. In Section 3.3, we present the comparison of three methods and hope it will benefit future research on similar problems.

3.2 Basic Framework

Our framework is a privacy-preserving action recognition pipeline that uses video data. It is a prototype of the in-demand privacy protection in smart camera applications. Figure 1 depicts the basic framework implementing the proposed formulation (3). The framework consists of three parts: the anonymization model f_A, the target model f_T, and the budget model f_B. f_A takes raw video $\mathcal{X}$ as input, filters out private information in $\mathcal{X}$, and outputs the anonymized video $f_A(\mathcal{X})$. f_T takes $f_A(\mathcal{X})$ as input and carries out the target utility task. f_B also take $f_A(\mathcal{X})$ as input and try to predict the private information from $f_A(\mathcal{X})$. All three models are implemented with deep neural networks, and their parameters are learnable during the training procedure. The entire pipeline is trained under the guidance of the hybrid loss of L_T and J_B. The training procedure has two goals. The first goal is to find an anonymization model f_A^* that can filter out the private information in the original video while keeping useful information for the target utility task. The second goal is to find a target model that can achieve good performance on the target utility task using anonymized videos $f_A(\mathcal{X})$. Similar frameworks have been used in feature disentanglement [48]–[51]. After training, the learned anonymization model can be applied on a local device (*e.g.*, smart camera), by designing an embedded chipset responsible for the anonymization at the hardware-level [37]. We can convert raw video to anonymized video locally and only transfer the anonymized video through the Internet to the backend (*e.g.*, cloud) for target utility task analysis. The private information in the raw videos will be unavailable on the backend.

Specifically, f_A is implemented using the model in [52], which can be taken as a 2D convolution-based frame-level filter. In other words, f_A converts each frame in $\mathcal{X}$ into a feature map of the same shape as the original frame. We use state-of-the-art human action recognition model C3D [53] as f_T and state-of-the-art image classification models, such as ResNet [54] and MobileNet [55], as f_B. Since the action recognition model we use is C3D, we need to split the videos into clips with a fixed frame number. Each clip

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3026709, IEEE Transactions on Pattern Analysis and Machine Intelligence

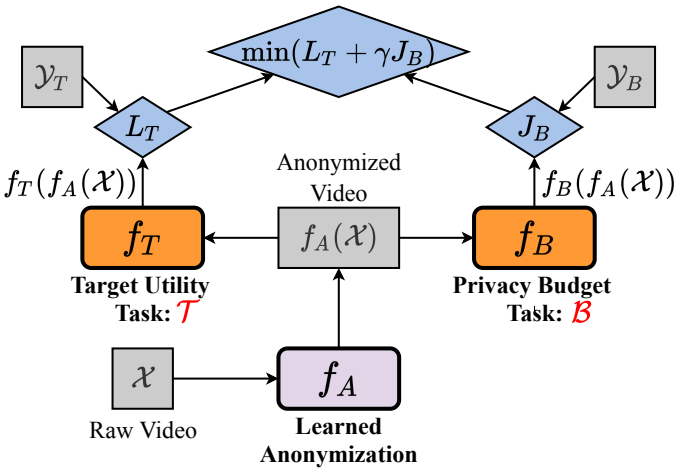


Fig. 1: Basic adversarial training framework for privacy-preserving action recognition.

is a 4D tensor of shape $[T, W, H, C]$, where T is the number of frames in each clip and W, H, C are the width, height, and the number of color channels in each frame respectively. Unlike f_T, which takes a 4D tensor as an input data sample, f_B takes a 3D tensor (*i.e.* a frame) as input. We average[1] the logits over the temporal dimension of each video clip to calculate J_B and predict the budget task label.

3.3 Optimization Strategies

In the following algorithms, we denote θ_B as the parameters of f_B. Similarly, f_A and f_T are parameterized by θ_A and θ_T respectively. $\alpha_A, \alpha_B, \alpha_T$ are learning rates used to update $\theta_A, \theta_B, \theta_T$. th_B and th_T are accuracy thresholds for the target utility task and the privacy budget prediction. max_iter is the maximum number of iterations. For simplicity concern, we abbreviate $L_T(f_T(f_A(\mathcal{X})), \mathcal{Y}_T)$, $J_B(f_B(f_A(\mathcal{X})), \mathcal{Y}_B)$, and $L_B(f_B(f_A(\mathcal{X})), \mathcal{Y}_B)$ as $L_T(\theta_A, \theta_T)$, $J_B(\theta_A, \theta_B)$[2] and $L_B(\theta_A, \theta_B)$ respectively. Acc is a function to compute accuracy on the privacy budget and the target utility tasks, given training data ($\mathcal{X}^t$, $\mathcal{Y}_B^t$ and $\mathcal{Y}_T^t$) and validation data ($\mathcal{X}^v$, $\mathcal{Y}_B^v$ and $\mathcal{Y}_T^v$).

3.3.1 Gradient reverse layer (GRL)

We consider Eq. (3) as a minimax problem [56]:

$$\theta_A^*, \theta_T^* = \text{argmin}_{(\theta_A, \theta_T)} L(\theta_A, \theta_T, \theta_B^*)$$
$$\theta_B^* = \text{argmax}_{\theta_B} L(\theta_A^*, \theta_T^*, \theta_B)$$

where $L(\theta_A, \theta_T, \theta_B) = L_T(\theta_A, \theta_T) + \gamma J_B(\theta_A, \theta_B) = L_T(\theta_A, \theta_T) - \gamma L_B(\theta_A, \theta_B)$.

GRL [47] is a state-of-the-art algorithm to solve such a minimax problem. The underlying mathematical gist is to solve the problem by alternative minimization:

$$\theta_A \leftarrow \theta_A - \alpha_A \nabla_{\theta_A} (L_T(\theta_A, \theta_T) - \gamma L_B(\theta_A, \theta_B)) \quad (4a)$$
$$\theta_T \leftarrow \theta_T - \alpha_T \nabla_{\theta_T} L_T(\theta_A, \theta_T) \quad (4b)$$
$$\theta_B \leftarrow \theta_B - \alpha_B \nabla_{\theta_B} L_B(\theta_A, \theta_B) \quad (4c)$$

1. AVERAGING the logits temporally gave a better performance in privacy budget prediction of J_B, compared with MAXING the logits.
2. Remember that J_B is the negative cross-entropy by definition.

We denote this method as **GRL** in the following parts and give the details in Algorithm 1.

Algorithm 1: GRL algorithm

```
Initialize θ_A, θ_T and θ_B;
for t ← 1 to max_iter do
    Update θ_A using Eq. (4a)
    while Acc(X^v, Y_T^v) ≤ th_T do
        Update θ_T using Eq. (4b)
    end
    while Acc(X^t, Y_B^t) ≤ th_B do
        Update θ_B using Eq. (4c)
    end
end
```

3.3.2 Alternating optimization of two loss functions

The goal in Eq. (3) can also be formulated as alternatively solving the following two optimization problems:

$$\theta_A^*, \theta_T^* = \text{argmin}_{(\theta_A, \theta_T)} L_T(\theta_A, \theta_T) \quad (5a)$$
$$\theta_B^*, \theta_A^* = \text{argmin}_{\theta_B} \text{argmax}_{\theta_A} L_B(\theta_A, \theta_B) \quad (5b)$$

Eq. (5a) is a standard minimization problem which can be solved by end-to-end training f_A and f_T. Eq. (5b) is a minimax problem which we solve by state-of-the-art minimax optimization method "K-Beam" [46]. K-Beam method keeps track of K different sets of budget model parameters (denoted as $\{\theta_B^i\}_{i=1}^K$) during training time, and alternatively updates θ_A and $\{\theta_B^i\}_{i=1}^K$.

Inspired by K-Beam method, we present the following parameter update rules to alternatively solve the two loss functions in Eq. (5):

$$\theta_A, \theta_T \leftarrow \theta_A, \theta_T - \alpha_T \nabla_{(\theta_T, \theta_A)} L_T(\theta_A, \theta_T) \quad (6a)$$
$$j \leftarrow \text{argmin}_{i \in \{1, \ldots, K\}} L_B(\theta_A, \theta_B^i) \quad (6b\text{-A})$$
$$\theta_A \leftarrow \theta_A + \alpha_A \nabla_{\theta_A} L_B(\theta_A, \theta_B^j) \quad (6b\text{-B})$$
$$\theta_B^i \leftarrow \theta_B^i - \alpha_B \nabla_{\theta_B^i} L_B(\theta_A, \theta_B^i), \ \forall i \in \{1, \ldots, K\} \quad (6c)$$

We denote this method as **Ours-K-Beam** in the following parts and give the details in Algorithm 2, where d_iter is the number of iterations used in the step of maximizing L_B.

3.3.3 Maximize entropy

We empirically find that minimizing negative cross-entropy J_B, which is a *concave* function, causes numerical instabilities in Eq. (4a). So, we replace J_B with $-H_B$, the negative entropy of $f_B(f_A(\mathcal{X}))$, which is a *convex* function[3]:

$$H_B(f_B(f_A(\mathcal{X}))) \triangleq H(f_B(f_A(\mathcal{X}))),$$

where $H(\cdot)$ is the entropy function. Minimizing $-H_B$ is equivalent to maximizing entropy, which will encourage "uncertain" predictions. We replace J_B in Eq. (4a) by $-H_B$,

3. This point discusses the convexity or concavity of different loss functions when viewing them as the outermost function in the composite function. Both loss functions are neither convex nor concave *w.r.t.* model weights.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3026709, IEEE Transactions on Pattern Analysis and Machine Intelligence

Algorithm 2: Ours-K-Beam algorithm

Initialize θ_A, θ_T and $\{\theta_B^i\}_{i=1}^K$;
for $t \leftarrow 1$ **to** *max_iter* **do**
  /*L_T **step:***/
  while $Acc(\mathcal{X}^v, \mathcal{Y}_T^v) \leq th_T$ **do**
    Update θ_T, θ_A using Eq. (6a)
  end
  /*L_B ***Max*** **step:***/
  Update j using Eq. (6b-A)
  for $t_2 \leftarrow 1$ **to** *d_iter* **do**
    Update θ_A using Eq. (6b-B)
  end
  /*L_B ***Min*** **step:***/
  for $i \leftarrow 1$ **to** K **do**
    while $Acc(\mathcal{X}^t, \mathcal{Y}_B^t) \leq th_B$ **do**
      Update θ_B^i using Eq. (6c)
    end
  end
end

abbreviate $H_B(f_B(f_A(\mathcal{X})))$ as $H_B(\theta_A, \theta_B)$, and propose the following new update scheme:

$$\theta_A \leftarrow \theta_A - \alpha_A \nabla_{\theta_A}(L_T(\theta_A, \theta_T) - \gamma H_B(\theta_A, \theta_B)) \quad (7a)$$

$$\theta_T, \theta_A \leftarrow \theta_T, \theta_A - \alpha_T \nabla_{\theta_T, \theta_A} L_T(\theta_A, \theta_T) \quad (7b)$$

$$\theta_B \leftarrow \theta_B - \alpha_B \nabla_{\theta_B} L_B(\theta_A, \theta_B) \quad (7c)$$

where L_T and L_B are still cross-entropy loss functions as in Eq. (4). Unlike in Eq. (4b), where we only update θ_T when minimizing L_T, we train θ_T and θ_A in an end-to-end manner as shown in Eq. (7b), since we find it achieves better performance in practice. We denote this method as **Ours-Entropy** in the following parts and give the details in Algorithm 3.

Algorithm 3: Ours-Entropy algorithm

Initialize θ_A, θ_T and θ_B;
for $t \leftarrow 1$ **to** *max_iter* **do**
  Update θ_A using Eq. (7a)
  while $Acc(\mathcal{X}^v, \mathcal{Y}_T^v) \leq th_T$ **do**
    Update θ_T, θ_A using Eq. (7b)
  end
  while $Acc(\mathcal{X}^t, \mathcal{Y}_B^t) \leq th_B$ **do**
    Update θ_B using Eq. (7c)
  end
end

3.4 Addressing the Dataset Challenge by Cross-Dataset Training and Evaluation: An Initial Attempt

An ideal dataset to train and evaluate our framework would be a set of human action videos with both action labels and privacy attributes provided. On the SBU dataset, we can use the actor pair as a simple privacy attribute. But when we want to evaluate our method on more complex privacy attributes, we run into the dataset challenge: To the best of our knowledge, no existing datasets have both human action labels and complex privacy attributes provided on the same videos.

Given the observation that a privacy attributes predictor trained on VISPR can correctly identify privacy attributes occurring in UCF101 and HMDB51 videos (examples in the Appendix C), we hypothesize that the privacy attributes have good "transferability" across UCF101/HMDB51 and VISPR. Therefore, we can use a privacy prediction model trained on VISPR to assess the privacy leak risk on UCF101/HMDB51.

In view of that, we propose to use cross-dataset training and evaluation as a workaround method. In brief, we train action recognition (target utility task) on human action datasets, such as UCF101 [57] and HMDB51 [58], and train privacy protection (budget task) on visual privacy dataset VISPR [13], while letting the two interact via their shared component - the learned anonymization model. More specifically, during training, we have two pipelines: one is f_A and f_T trained on UCF101 or HMDB51 for action recognition; the other is f_A and f_B trained on VISPR to suppress multiple privacy attribute prediction. The two pipelines share the same parameters for f_A. During the evaluation, we evaluate model utility (*i.e.*, action recognition) on the testing set of UCF101 or HMDB51 and privacy protection performance on the testing set of VISPR. Such cross-dataset training and evaluation shed new possibilities on training privacy-preserving recognition models, even under the practical shortages of datasets that have been annotated for both tasks. Notably, "cross-dataset training" and "cross-dataset testing (or evaluation)" are two independent strategies used in this paper; they can be used either together or separately. Details of our three experiments (SBU, UCF-101, and HMDB51) are explained as follows:

- SBU (Section 4.1): we train and evaluate our framework on the same video set, by considering actor identity as a simple privacy attribute. *Neither cross-training nor cross-evaluation is involved*.
- UCF101 (Section 4.2): we perform *both cross-training and cross-evaluation*, on UCF-101 + VISPR. Such a method provides an alternative to flexibly train and test privacy-preserving video recognition for different utility/privacy combinations, without annotating specific datasets.
- HMDB51 (Section 5.5), we use cross-training on HMDB51 + VISPR datasets similarly to the UCF-101 experiment; but for testing, we evaluate both utility and privacy performance on the same, newly-annotated PA-HMDB51 testing set. Therefore, it involves *only cross-training, but not cross-evaluation*.

Beyond the above initial attempt, we further construct a new dataset dedicated to the privacy-preserving action recognition task, which will be presented in Section 5.

3.5 Addressing the $\forall$ Challenge by Privacy Budget Model Restarting and Ensemble

To improve the generalization ability of learned f_A over all possible $f_B \in \mathcal{P}$ (*i.e.*, privacy cannot be reliably predicted by any model), we hereby discuss two simple and easy-to-implement options. Other more sophisticated model re-sampling or model search approaches, such as [59], will be explored in future work.

3.5.1 Privacy Budget Model Restarting

Motivation The max step over $J_B(f_B(f_A(\mathcal{X})), \mathcal{Y}_B)$ in Eq. (3) leads to the optimizer being stuck in bad local solutions (similar to "mode collapse" in GANs), that will hurdle the entire minimax optimization. Model restarting provides a mechanism to "bypass" the bad solution when it occurs, thus enabling the minimax optimizer to explore better solution.

Approach At certain point of training (*e.g.*, when the privacy budget $L_B(f_B(f_A(\mathcal{X})), \mathcal{Y}_B)$ stops decreasing any further), we re-initialize f_B with random weights. Such a random restarting aims to avoid trivial overfitting between f_B and f_A (*i.e.*, f_A is only specialized at confusing the current f_B), without requiring more parameters. We then start to train the new model f_B to be a strong competitor, *w.r.t.* the current $f_A(\mathcal{X})$: specifically, we freeze the training of f_A and f_T, and change to minimizing $L_B(f_B(f_A(\mathcal{X})), \mathcal{Y}_B)$, until the new f_B has been trained from scratch to become a strong privacy prediction model over current $f_A(\mathcal{X})$. We then resume adversarial training by unfreezing f_A and f_T, as well as switching the loss for f_B back to the adversarial loss (negative entropy or negative cross-entropy). Such a random restarting can repeat multiple times.

3.5.2 Privacy Budget Model Ensemble

Motivation Ideally in Eq. (3) we should maximize the error over the "current strongest possible" attacker f_A from $\mathcal{P}$ (a large and continuous f_B family), over which searching/sampling is impractical. Therefore we propose a privacy budget model ensemble as an approximation strategy, where we approximate the continuous $\mathcal{P}$ with a discrete set of M sample functions. Such a strategy is empirically verified in Section 4 and 5 to address the critical "$\forall$ Challenge" in privacy protection, *i.e.*, enhancing the defense against unseen attacker models (compared to the clear "attacker overfitting" phenomenon when sticking to one f_A during training).

Approach Given the budget model ensemble $\overline{\mathcal{P}}_t \triangleq \{f_B^i\}_{i=1}^M$, where M is the number of f_Bs in the ensemble during training, we turn to minimize the following discretized surrogate of Eq. (2):

$$f_A^*, f_T^* = \text{argmin}_{f_A, f_T}[L_T(f_T(f_A(\mathcal{X})), \mathcal{Y}_T) + \gamma \max_{f_B^i \in \overline{\mathcal{P}}_t} J_B(f_B^i(f_A(\mathcal{X})), \mathcal{Y}_B)] \quad (8)$$

The previous basic framework is a special case of Eq. (8) with $M = 1$. The ensemble strategy can be easily combined with restarting.

3.5.3 Combine Budget Model Restarting and Budget Model Ensemble with Ours-Entropy

Budget Model Restarting and Budget Model Ensemble can be easily combined with all three optimization schemes described in Section 3.3. We take Ours-Entropy as an example here. When model ensemble is used, we abbreviate $L_B(f_B^i(f_A(\mathcal{X})), \mathcal{Y}_B)$ and $H_B(f_B^i(f_A(\mathcal{X})))$ as $L_B(\theta_A, \theta_B^i)$ and $H_B(\theta_A, \theta_B^i)$ respectively. The new parameter updating scheme is:

$$\theta_A \leftarrow \theta_A - \alpha_A \nabla_{\theta_A}(L_T + \gamma \max_{\theta_B^i \in \overline{\mathcal{P}}_t} - H_B(\theta_A, \theta_B^i)) \quad (9a)$$

$$\theta_A, \theta_T \leftarrow \theta_A, \theta_T - \alpha_T \nabla_{(\theta_A, \theta_T)} L_T(\theta_A, \theta_T) \quad (9b)$$

$$\theta_B^i \leftarrow \theta_B^i - \alpha_B \nabla_{\theta_B^i} L_B(\theta_A, \theta_B^i),\ \forall i \in \{1, \ldots, M\} \quad (9c)$$

That's to say, we only suppress the model f_B^i with the largest privacy leakage $-H_B$, *i.e.*, the "most confident" one about its current privacy prediction, when updating anonymization model. But we still update all M budget models on the budget task. The formal description of Ours-Entropy with model restarting and ensemble is given in Algorithm 4, where $\{\theta_B^i\}_{i=1}^M$ is reinitialized every $rstrt\_iter$ iterations. Likewise, GRL and Our-K-Beam can also be combined with restarting and ensemble, whoses details are shown in Appendix A.

Algorithm 4: Ours-Entropy algorithm (with model restarting and model ensemble)

```
Initialize θ_A, θ_T and θ_B;
for t ← 1 to max_iter do
    if t ≡ 0 (mod rstrt_iter) then
        Reinitialize {θ_B^i}_{i=1}^M
    end
    Update θ_A using Eq. (9a)
    while Acc(X^v, Y_T^v) ≤ th_T do
        Update θ_T, θ_A using Eq. (9b)
    end
    for i ← 1 to M do
        while Acc(X^t, Y_B^t) ≤ th_B do
            Update θ_B^i using Eq. (9c)
        end
    end
end
```

3.6 Two-Step Evaluation Protocol

The solution to Eq. (2) gives an anonymization model f_A^* and a target utility task model f_T^*. We need to evaluate f_A^* and f_T^* on the trade-off they have achieved between target task utility and privacy protection in two steps: (1) whether the learned target utility task model maintains satisfactory performance on anonymized videos; (2) whether the performance of an arbitrary privacy prediction model on anonymized videos will deteriorate.

Suppose we have a training dataset X^t with target and budget task ground truth labels $\mathcal{Y}_T^t$ and $\mathcal{Y}_B^t$, and an evaluation dataset X^e with target and budget task ground truth labels $\mathcal{Y}_T^e$ and $\mathcal{Y}_B^e$. In the first step, when evaluating the target task utility, we should follow the traditional routine: compare $f_T^*(f_A^*(X^e))$ with $\mathcal{Y}_T^e$ to get the evaluation accuracy on the target utility task, denoted as A_T, which we expect to be as high as possible. In the second step, when evaluating the privacy protection, it is insufficient if we only observe that the learned f_A^* and f_B^* lead to poor classification accuracy on X^e, because of the $\forall$ challenge: the attacker can select any privacy budget model to steal private information from anonymized videos $f_A^*(X^e)$. To empirically verify that

f_A^* prohibits reliable privacy prediction for other possible budget models, we propose a novel procedure:

- We randomly re-sample N privacy budget prediction models $\overline{\mathcal{P}}_e \triangleq \{f_B^i\}_{i=1}^N$ from $\mathcal{P}$ for evaluation. Note that these N models used in evaluation $\overline{\mathcal{P}}_e$ have no overlap with the M privacy budget model ensemble $\overline{\mathcal{P}}_t$ used in training (*i.e.*, $\overline{\mathcal{P}}_e \cap \overline{\mathcal{P}}_t = \varnothing$).
- We train these N models $\overline{\mathcal{P}}_e$ on anonymized training videos $f_A^*(X^t)$ to make correct predictions on private information, *i.e.*, $\min_{f_B^i} L_B(f_B^i(f_A^*(X^t)), \mathcal{Y}_B^t)$ for all $i \in \{1, \ldots, N\}$. Note that f_A^* is fixed during this training procedure.
- After that, we apply each f_B^i on anonymized evaluation videos $f_A^*(X^e)$ and compare the outputs $f_B^i(f_A^*(X^e))$ with $\mathcal{Y}_B^e$ to get privacy budget accuracy of the i-th budget model.
- We select the highest accuracy among all N privacy budget models and use it as the final privacy budget accuracy A_B^N, which we expect to be as low as possible.

4 Simulation Experiments

We show the effectiveness of our framework on *privacy-preserving action recognition* on existing datasets.

Overview of Experiment Settings The target utility task is human action recognition, since it is a highly demanded feature in smart home and smart workplace applications. Experiments are carried out on three widely used human action recognition datasets: SBU Kinect Interaction Dataset [60], UCF101 [57] and HMDB51 [58]. The privacy budget task varies in different settings. In the SBU dataset experiments, the privacy budget is to prevent the videos from leaking human identity information. In the experiments on UCF101 and HMDB51, the privacy budget is to protect visual privacy attributes as defined in [13]. We emphasize that the general framework proposed in Section 3.2 can be used for a large variety of target utility tasks and privacy budget task combinations, not only limited to the aforementioned settings.

Following the notations in Section 3.2, on all the video action recognition datasets including SBU, UCF101 and HMDB51, we set $W = 112$, $H = 112$, $C = 3$, and we set $T = 16$ (C3D's required temporal length and spatial resolution). Note that the original resolution for SBU, UCF101 and HDMB51 are 640×480, 320×240 and 320×240, respectively. We downsample video frames to resolution 160×120. To reduce the spatial resolution to 112×112, we use random-crop and center-crop in training and evaluation, respectively.

Baseline Approaches We consistently use two groups of approaches as baselines across the three action recognition datasets. These two groups of baselines are *naive downsamples* and *empirical obfuscations*. The group of *naive downsamples* chooses downsample rates from $\{1, 2, 4, 8, 16\}$, where 1 stands for no down-sampling. The group of *empirical obfuscations* includes approaches selected from different combinations in {box, segmentation} $\times$ {blurring, blackening} $\times$ {face, human body}. Details are listed below:

- *Naive Downsamle*: Spatially downsample each frame.
- *Box-Black-Face*: Boxing and blackening faces.
- *Box-Black-Body*: Boxing and blackening bodies.
- *Seg-Black-Face*: Segmenting and blackening faces.
- *Seg-Black-Body*: Segmenting and blackening bodies.
- *Box-Blur-Face*: Boxing and blurring faces.
- *Box-Blur-Body*: Boxing and blurring bodies.
- *Seg-Blur-Face*: Segmenting and blurring faces.
- *Seg-Blur-Body*: Segmenting and blurring bodies.

Our Proposed Approaches The previous two groups of baselines are compared with our proposed three approaches:

- *GRL*: as described in Section 3.3.1.
- *Ours-K-Beam*: as described in Section 3.3.2. We have tried $K = 1, 2, 4, 8$.
- *Ours-Entropy*: as described in Section 3.3.3. In the privacy budget model ensemble $\overline{\mathcal{P}}_t$, the M models are chosen from MobileNet-V2 [61] family with different width multipliers. We have tried $M = 1, 2, 4, 8$.

All three approaches are evaluated with and without privacy budget model restarting.

Evaluation In the two-step evaluation (as described in Section 3.6), we have used $N = 10$ different state-of-the-art classification networks, namely ResNet-V1-{50,101} [54], ResNet-V2-{50,101} [62], Inception-V1 [63], Inception-V2 [64], and MobileNet-V1-{0.25,0.5,0.75,1} [55], as $\overline{\mathcal{P}}_e$. Note that $\overline{\mathcal{P}}_e \cap \overline{\mathcal{P}}_t = \varnothing$. All detailed numerical results reported in following sections can be found in Appendix B.

4.1 Identity-Preserving Action Recognition on SBU: Single-Dataset Training

We compare our proposed approaches with the groups of baseline approaches to show our methods' significant superiority in balancing privacy protection and model utility. We use three different optimization schemes described in Section 3.3 on our framework and empirically show all three largely outperform the baseline methods. We also show that adding the model ensemble and model restarting, as described in Section 3.5, to the optimization procedure can further improve the performance of our method.

4.1.1 Experiment Setting

SBU Kinect Interaction Dataset [60] is a two-person interaction dataset for video-based action recognition. 7 participants performed actions, and the dataset is composed of 21 sets. Each set uses different pairs of actors to perform all 8 interactions. However, some sets use the same two actors but with different actors acting and reacting. For example, in set 1, actor 1 is acting, and actor 2 is reacting; in set 4, actor 2 is acting, and actor 1 is reacting. These two sets have the same actors, so we combine them as one class to better fit our experimental setting. In this way, we combine all sets with the same actors and finally get 13 different actor pairs. This dataset's target utility task is action recognition, which could be taken as a classification task with 8 different classes. The privacy budget task is to recognize the actor pairs of the videos, which could be taken as a classification task with 13 different classes.

4.1.2 Implementation Details

In Algorithms 1-3, we set step sizes $\alpha_T = 10^{-5}$, $\alpha_B = 10^{-2}$, $\alpha_A = 10^{-4}$, accuracy thresholds $th_T = 85\%$, $th_B = 99\%$ and $max\_iter = 800$. In Algorithm 2, we set d_iter to

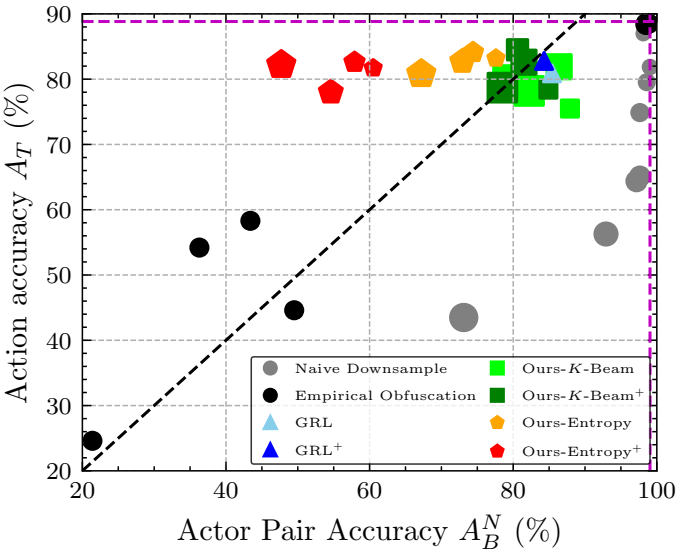


Fig. 2: Privacy-utility trade-off on SBU dataset. For *Naive Downsample* method, a larger marker means a larger adopted down-sampling rate. For *Ours-K-Beam* method, a larger marker means a larger K (number of beams) in Algorithm 2. For *Ours-Entropy* and *Ours-Entropy (restarting)*, a larger marker means a larger M (number of ensemble models) in Algorithm 4. Methods with "+" superscript are combined with model restarting. Vertical and horizontal purple dashed lines indicate A_B^N and A_T on the original non-anonymized videos, respectively. The black dashed line indicates where $A_B^N = A_T$. Detailed experimental settings and numerical results for each method can be found in Appendix B.

be 30. In Algorithm 4, we set $rstrt\_iter$ to be 100. Other hyper-parameters of Algorithm 4 are identical with those in Algorithm 3. We set γ=2 in loss function (3) and use Adam optimizer [65] to update all parameters.

4.1.3 Results and Analyses

We present the experimental results of our proposed methods and other baseline methods in Figure 2, which shows the trade-off between the action recognition accuracy A_T, and the actor pair recognition accuracy A_B^N. In order to interpret this figure, we should note that a desirable trade-off should incur maximal target accuracy A_T (y-axis) and minimal privacy budget accuracy A_B^N (x-axis). Therefore, a point closer to the top-left corner represents an anonymization model f_A^* with more desirable performance. The magenta dotted line suggests the target accuracy A_T on original unprotected videos. This can be roughly considered as the A_T upper bound for all privacy protection methods, under the assumption that f_A^* will unavoidably filter out some useful information for the target utility task.

As we can see, Ours-K-Beams, Ours-Entropy, and GRL all largely outperform the two groups of naive baselines. {box, segmentation} $\times$ {blurring, blackening} $\times$ {face} and naive downsample with a low rate (*e.g.* 2 and 4) can lead to decent action accuracy, but the privacy budget accuracy A_B^N is still very high, meaning these methods fail to protect privacy. On the other hand, {box, segmentation} $\times$ {blurring, blackening} $\times$ {body} and naive downsample with a high rate (*e.g.* 8 and 16) can effectively suppress A_B^N to a low level, but A_T also suffers a huge negative impact, which means the anonymized videos are of little practical utility. Our methods, in contrast, achieve a great balance between utility and privacy protection. Ours-Entropy can decrease A_B^N by around 30% with nearly no harm on A_T.

Comparison of three methods K-Beam is a state-of-the-art minimax optimization problem, and we apply it to solve a sub-problem (*i.e.*, Eq. (5b)) of our more complex three-party competition game. Unfortunately, we empirically find the K-Beam algorithm becomes more unstable when we introduce a new competing party to the minimax optimization problem. GRL is originally proposed for domain adaptation. On our new visual privacy protection task, we find it unstable and sensitive to model initialization. By replacing the *concave* negative cross-entropy loss function with the *convex* negative entropy, Ours-Entropy empirically stabilizes the optimization and gives the best performance among all three methods.

The results also show the effectiveness of model restarting and model ensemble: model restarting can suppress A_B^N much further, with no additional harm on A_T, and model ensemble with larger M also tends to improve the trade-off.

4.2 Action Recognition on UCF101 with Multiple Privacy Attributes Protected: Cross-Dataset Training and Evaluation

4.2.1 Experiment Setting

UCF101 is an action recognition dataset with 13,320 real-life human action videos collected from YouTube. It contains videos of 101 different actions. We use the official train-test split for this dataset. The target utility task $\mathcal{T}$ is to do human action recognition on UCF101, which can be taken as a video classification task with 101 classes.

VISPR is a dataset with 22,167 images annotated with 68 privacy attributes, *e.g.*, semi-nudity, hobbies, face, race, gender, skin color, and so on. Each attribute of an image is labeled as "present" or "non-present" depending on whether the specific privacy attribute information is contained in the image. Among the 68 attributes, there are 7 attributes that frequently appear in both UCF101 datasets and the smart home videos. Therefore we select these 7 attributes for protection in our experiments. These 7 attributes are semi-nudity, occupation, hobbies, sports, personal relationship, social relationship, and safe. The privacy budget task $\mathcal{B}$ is to predict privacy attributes on the VISPR dataset, which can be taken as a multi-label image classification task (7 labels, each is a binary classification task). We define A_B^N as class-based mean average precision (cMAP) [13] to measure the performance of this privacy budget task. The official train-test split is used on the VISPR dataset.

4.2.2 Implementation Details

In Algorithm 4, we set step sizes $\alpha_T = 10^{-5}$, $\alpha_B = 10^{-2}$, $\alpha_A = 10^{-4}$, accuracy thresholds $th_T = 70\%$, $th_B = 99\%$, $max\_iter = 800$ and $rstrt\_iter = 100$. We set $\gamma = 0.5$ in Eq. (3) and use Adam optimizer to update all parameters. Values of K and M are identical to those in SBU experiments.

4.2.3 Results and Analyses

We present the experimental results in Figure 3. All naive downsample and empirical obfuscation methods cause A_T to drop dramatically while A_B^N only drops a little bit, which means the utility of videos is greatly reduced while the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3026709, IEEE Transactions on Pattern Analysis and Machine Intelligence

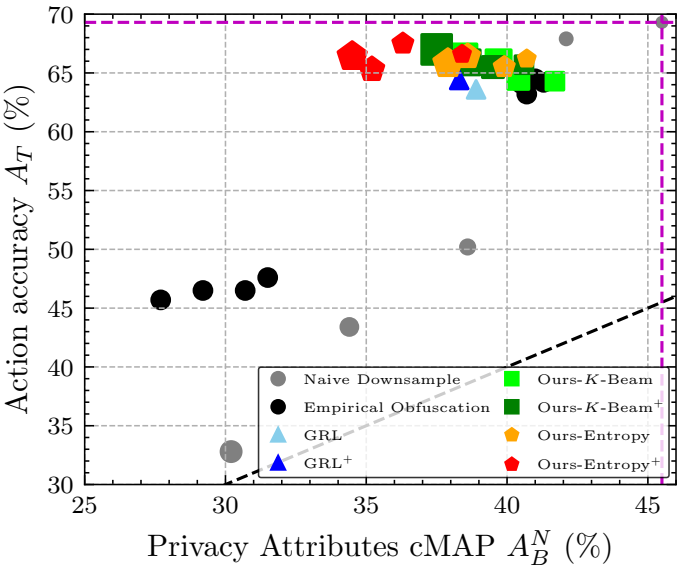


Fig. 3: Privacy-utility trade-off on UCF-101/VISPR Dataset. For *Naive Downsample* method, a larger marker means a larger down sampling rate is adopted. For *Ours-K-Beam* method, a larger marker means a larger K (number of beams) in Algorithm 2. For *Ours-Entropy* and *Ours-Entropy (restarting)*, a larger marker means a larger M (number of ensemble models) in Algorithm 4. Methods with "+" superscript are combined with model restarting. Vertical and horizontal purple dashed lines indicate A_B^N and A_T on the original non-anonymized videos, respectively. The black dashed line indicates where $A_B^N = A_T$. Detailed experimental settings and numerical results for each method can be found in Appendix B.

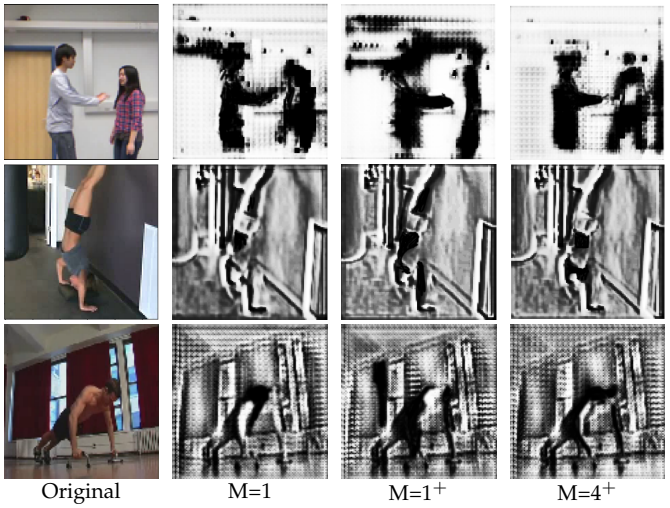


Fig. 4: The center frame of example videos before (column 1) and after (columns 2-4) applying the anonymization transform learned by Ours-Entropy. The first row shows a frame from a "pushing" video in the SBU dataset; the second row shows a frame from a "handstand" video in the UCF101 dataset; the third row shows a frame from a "push-up" video in the PA-HMDB51 dataset. Privacy attributes in the last two rows include semi-nudity, face, gender, and skin color. Model restarting and ensemble settings are indicated below each anonymized image. M is the number of ensemble models. Methods with a "+" superscript are combined with model restarting.

private information is hardly filtered out. In contrast, with the help of model restarting and model ensemble, Ours-Entropy can decrease A_B^N by 7% while keeping A_T as high as that on the original raw videos, meaning the privacy is protected at almost no cost on the utility. Hence, Ours-Entropy outperforms all naive downsample and empirical obfuscation baselines in this experiment. It also shows an advantage over GRL and Ours-K-Beam.

4.3 Anonymized Video Visualization

We provide the visualization of the anonymized videos on SBU, UCF101, and our new dataset PA-HMDB51 (see Section 5) in Figure 4. To save space, we only show the center frame of each anonymized video. The visualization shows that the privacy attributes in the anonymized videos are filtered out, but it is still possible to recognize the actions.

5 PA-HMDB51: A New Benchmark

5.1 Motivation

There is no public dataset containing both human action and privacy attribute labels on the same videos to the best of our knowledge. This poses two challenges. Firstly, the lack of available datasets has increased the difficulty in employing a data-driven joint training method. Secondly, this complication has made it impossible to directly evaluate the privacy-utility trade-off achieved by a learned anonymization model f_A^*. To solve this problem, we annotate and present the very first human action video dataset with privacy attributes labeled, named **PA-HMDB51** (Privacy Annotated HMDB51). We evaluate our method on this newly built dataset and further demonstrate our method's effectiveness.

5.2 Selecting and Labeling Privacy Attributes

A recent work [13] has defined 68 privacy attributes which could be disclosed by images. However, most of them seldom make any appearance in public human action datasets. We carefully select 7 privacy attributes that are most relevant to our smart home settings out of the 68 attributes from [13]. The 7 attributes we use are skin color, gender, face (partial), face (complete), nudity, personal relationship, and social circle. We further combine those 7 attributes into 5 to better fit the human action videos: combining "face (partial)" and "face (complete)" into one attribute "face" and combining "personal relationship" (only intimate relationships such as friends, couples or family members are considered in our setting) and "social circle" (*e.g.* colleagues and classmates) into "relationship". To this end, we have 5 privacy attributes that widely appear in public human action datasets and are closely relevant to our smart home setting. The detailed description of each attribute, their possible ground truth values and their corresponding meanings are listed in Table 1. Some annotated frames in our PA-HMDB51 dataset are shown in Table 2 as examples.

Privacy attributes may vary during the video clip. For example, in some frames we may see a person's full face, while in the next frames the person may turn around and his/her face is no longer visible. Therefore, we decide to label all privacy attributes for each frame.[4]

The annotation of privacy labels was manually performed by a group of students at the CSE department of Texas A&M University. Each video was annotated by at least three individuals and then cross-checked.

4. A tiny portion of frames in some HMDB51 videos do not contain any person. No privacy attributes are annotated for those frames.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3026709, IEEE Transactions on Pattern Analysis and Machine Intelligence

5.3 HMBD51 as the Data Source

Now that we have defined the 5 privacy attributes, we need to identify a source of human action videos for annotation. There are a number of choices available, such as [57], [58], [66]–[68]. We choose HMDB51 [58] to label privacy attributes since it consists of more diverse private information, especially nudity/semi-nudity and relationship.

We provide a per-frame annotation of the selected 5 privacy attributes on 515 videos selected from HMDB51. In this paper, we treat all 515 videos as testing samples[5]; however, we do not exclude the future possibility of using them for training. Our ultimate goal would be to create a larger-scale version of PA-HMDB51 that allows for both training and testing coherently on the same benchmark. For now, we use PA-HMDB51 to facilitate better testing, while still considering cross-dataset training as a rough yet useful option to train privacy-preserving video recognition (before the larger dataset becomes available).

5.4 Dataset Statistics

5.4.1 Action Distribution

When selecting videos from the HMDB51 dataset, we consider two criteria on action labels. First, the action labels should be balanced. Second (and more implicitly), we select more videos with non-trivial privacy labels. For example, "brush hair" action contains many videos with a "semi-nudity" attribute and "pull-up" action contains many videos with a "partially visible face" attribute. Despite their practical importance, these privacy attributes are relatively less seen in the entire HMDB51 dataset, so we tend to select more videos with these attributes, regardless of their action classes. The resultant distribution of action labels is depicted in Figure 5 (left panel), showing a relative class balance.

5.4.2 Privacy Attribute Distribution

We try to make the label distribution for each privacy attribute as balanced as possible by manually selecting those videos containing uncommon privacy attribute values in original HMDB51 to label. For instance, videos with semi-nudity are overall uncommon, so we deliberately select those videos containing semi-nudity into our PA-HMDB51 dataset. Naturally, people are reluctant to release data that contains privacy concerns to the public, so the privacy attributes are highly unbalanced in any public video datasets. Although we have used this method to reduce the data imbalance, the PA-HMDB51 is still unbalanced. Frame-level label distributions of all 5 privacy attributes are shown in Figure 6.

5.4.3 Action-Attribute Correlation

If there is a strong correlation between a privacy attribute and an action, it would be harder to remove the private information from the videos without much harm to the action recognition task. For example, we would expect a high correlation between the attribute "gender" and the action "brush hair" since this action is carried out much more often by females than by males. We show the correlation between privacy attributes and actions in Figure 5 (right panel) and more details in Appendix D.

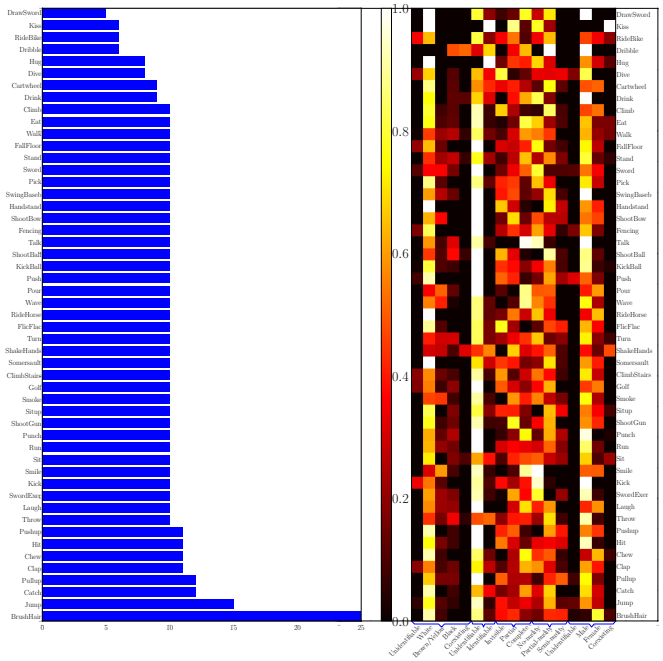


Fig. 5: **Left**: action distribution of PA-HMDB51. Each column shows the number of videos with a certain action. *E.g.*, the last bar shows there are 25 "brush hair" videos in the PA-HMDB51 dataset. **Right**: action-attribute correlation in the PA-HMDB51 dataset. The color represents ratio of the number of frames of some action containing a specific privacy attribute value *w.r.t.* the total number of frames of the action.

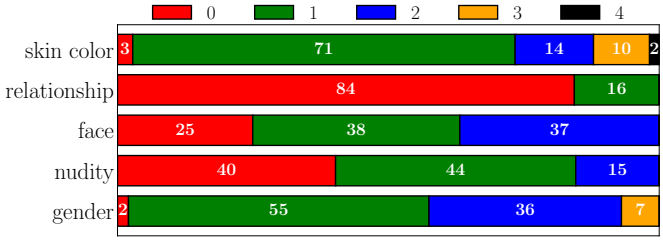


Fig. 6: Label distribution per privacy attribute in the PA-HMDB51. The rounded ratio numbers are shown as white text (in % scale). Definitions of label values $(0, 1, 2, 3, 4)$ for each attribute are described in Table 1.

5.5 Benchmark Results on PA-HMDB51: Cross-Dataset Training

5.5.1 Experiment Setting

We train our models using cross-dataset training on HMDB51 and VISPR datasets as we did in Section 4.2, except that we use the 5 attributes defined in Table 1 on VISPR

5. Labeling per-frame privacy attributes on a video dataset is extremely labor-consuming and subjective (needing individual labeling then cross-checking). As a result, the current size of PA-HMDB51 is limited. So far we have only used PA-HMDB51 as the testing set, and we seek to annotate more data and hopefully expand PA-HMDB51 for training as future work.

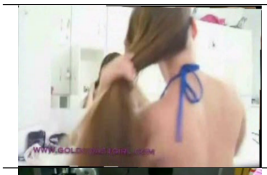
6. For "skin color" and "gender", we allow multiple labels to coexist. For example, if a frame showed a black person's shaking hands with a white person, we would label "black" and "white" for the "skin color" attribute. In the visualization, we use "coexisting" to represent the multi-label coexistence and we don't show in detail whether it is "white and black coexisting" or "black and yellow coexisting". For the rest three attributes, we label each attribute using the highest privacy-leakage risk among all persons in the frame. For example, given a frame where a group of people are hugging, if there is at least one complete face visible, we would label the "face" attribute as "completely visible".

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3026709, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1: Attribute Definition in the PA-HMDB51 Dataset

Attribute	Possible Values & Meaning
Skin Color	**0**: Skin color of the person(s) is/are unidentifiable. **1**: Skin color of the person(s) is/are white. **2**: Skin color of the person(s) is/are brown/yellow. **3**: Skin color of the person(s) is/are black. **4**: Persons with different skin colors are coexisting.[6]
Face	**0**: Invisible ($<10\%$ area is visible). **1**: Partially visible ($\geq 10\%$ but $\leq 70\%$ area is visible). **2**: Completely visible ($>70\%$ area is visible).
Gender	**0**: The gender(s) of the person(s) is/are unidentifiable. **1**: The person(s) is/are male. **2**: The person(s) is/are female. **3**: Persons with different genders are coexisting.
Nudity	**0**: No-nudity w/ long sleeves and pants **1**: Partial-nudity w/ short sleeves, skirts, or shorts **2**: Semi-nudity w/ half-naked body
Relationship	**0**: Relationship is unidentifiable. **1**: Relationship is identifiable.

TABLE 2: Example Annotated Frames in the PA-HMDB51 Dataset

Frame	Action	Privacy Attributes
	Brush hair	• skin color: white • face: invisible • gender: female • nudity: semi-nudity • relationship: unidentifiable
	Situp	• skin color: black • face: completely visible • gender: male • nudity: semi-nudity • relationship: unidentifiable

instead of the 7 used in Section 4.2. The trained models are directly evaluated on the PA-HMDB51 dataset[7] for both target utility task $\mathcal{T}$ and privacy budget task $\mathcal{B}$, without any re-training or adaptation. We exclude the videos in the PA-HMDB51 from the HMDB51 to get the training set. Similar to the UCF101 experiments, the target utility task $\mathcal{T}$ (*i.e.*, action recognition) can be taken as a video classification problem with 51 classes, and the privacy budget task $\mathcal{B}$ (*i.e.*, privacy attribute prediction) can be taken as a multi-label image classification task with two classes for each privacy attribute label. Notably, although PA-HMDB51 has provided concrete multi-class labels with specific privacy attribute classes, we convert them into binary labels during testing. For example, for "gender" attribute, we have provided ground truth labels "male", "female", "coexisting" and "cannot tell", but we only use "can tell" and "cannot tell" in our experiments, via combining "male", "female" and "coexisting" into the one class of "can tell". This is because we must keep the testing protocol on PA-HMDB51 consistent with the training protocol on VISPR (a multi-label, "either-or" type *binary* classification task, so that our models cross-trained on UCF101-VISPR can be evaluated directly. We hope to extend training to PA-HMDB51 in the future so that the privacy budget task can be formulated and evaluated as a multi-label, *multi*-classification problem.

All implementation details are identical with the UCF101 case, except that we adjust $th_T = 0.7$ and $th_B = 0.95$.

5.5.2 Results and Analysis

The results on PA-HMDB51 are shown in Figure 7. Our methods achieve much better privacy-utility trade-off compared with baseline methods. When $M = 4$, our methods can decrease privacy cMAP by around 8% with little harm to utility accuracy. Overall, the privacy gains are more limited compared to the previous two experiments, because no (re-)training is performed; but the overall comparison trends show the same consistency.

Asymmetrical Privacy Attributes Protection Cost Different privacy attributes have different protection costs. After applying the learned anonymization optimized by *Ours-Entropy (restarting, M=4)* on PA-HMDB51, the drop in AP of "face" is much more significant than "gender", which indicates that the "gender" attribute is much harder to suppress than "face". Such observation agrees that the gender attribute can be revealed by face, body, clothing, and even hairstyle. In future work, we will take such cost asymmetry into account by using a weighted loss combination of different privacy attributes or training dedicated privacy protector for the most informative private attribute.

Human Study on the Privacy Protection of Our Learned Anonymization We use human study to evaluate the privacy-utility trade-off achieved by our learned anonymization transform. We take both privacy protection and action recognition into account in the study. We emphasize here that both privacy protection and action recognition are evaluated on video level. There are 515 videos distributed on 51 actions in the PA-HMDB51. For each action in the PA-HMDB51, we randomly pick one video for human study. Among the 51 selected videos, we only keep 30 videos to reduce the human evaluation cost. There were 40 volunteers involved in the human study. In the study, they were asked to label all the privacy attributes and the action type on the raw videos and the anonymized videos. According to the experimental results (shown in Appendix E), the actions in the anonymized videos are still distinguishable to humans, but the privacy attributes are not recognizable at all. This human study further justifies that our learned anonymization transform can protect privacy and maintain target utility task performance simultaneously.

6 Conclusion

We propose an innovative framework to address the newly-established problem of privacy-preserving action recognition. To tackle the challenging adversarial learning process, we investigate three different optimization schemes. To further tackle the $\forall$ challenge of universal privacy protection, we propose the privacy budget model restarting and ensemble strategies. Both are shown to improve the privacy-utility trade-off further. Various simulations verified the effectiveness of the proposed framework. More importantly, we collect the first dataset for privacy-preserving video

7. We only use PA-HMDB51 as the testing set so far, since the current size of PA-HMDB51 is limited for training.

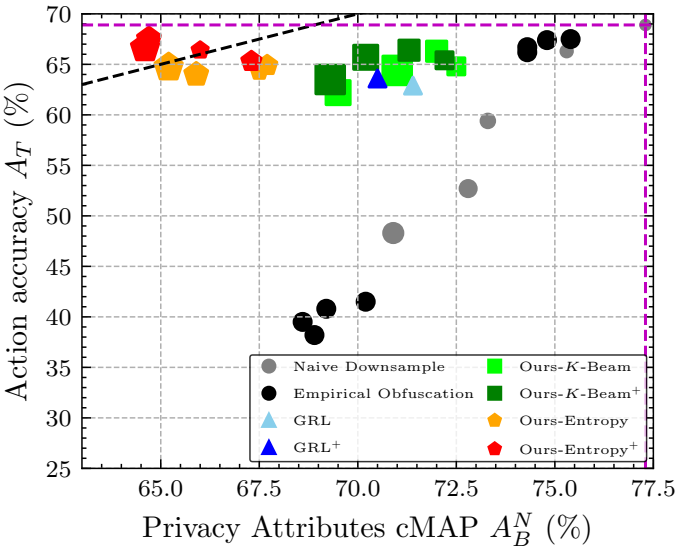


Fig. 7: Privacy-utility trade-off on PA-HMDB51 Dataset. For *Naive Downsample* method, a larger marker means a larger downsampling rate is adopted. For *Ours-K-Beam* method, a larger marker means a larger K (number of beams) in Algorithm 2. For *Ours-Entropy* and *Ours-Entropy (restarting)*, a larger marker means a larger M (number of ensemble models) in Algorithm 4. Methods with "+" superscript are combined with model restarting. Vertical and horizontal purple dashed lines indicate A_B^N and A_T on the original non-anonymized videos, respectively. The black dashed line indicates where $A_B^N = A_T$. Detailed experimental settings and numerical results for each method can be found in Appendix B.

action recognition, an effort that we hope could engage a broader community into this research field.

We note that there is much room to improve the proposed framework before it can be used in practice. For example, the definition of privacy leakage risk is core to the framework. Considering the $\forall$ challenge, current L_B defined with any specific f_B is insufficient; the privacy budget model ensemble could only be viewed as a rough discretized approximation of $\mathcal{P}$. More elaborated ways to model this $\forall$ challenge may lead to a further breakthrough in achieving the optimization goal.

Acknowledgement

Z. Wu, H. Wang, and Z. Wang were partially supported by NSF Award RI-1755701. The authors would like to sincerely thank Scott Hoang, James Ault, and Prateek Shroff for assisting the labeling of the PA-HMDB51 dataset.

References

[1] C. H. Donna Lu, "How abusers are exploiting smart home devices?" October 2019. [Online]. Available: https://www.vice.com/en_in/article/d3akpk/smart-home-technology-stalking-harassment

[2] M. W. Tobias, "Is your smart security camera protecting your home or spying on you?" August 2016. [Online]. Available: https://www.forbes.com/sites/marcwebertobias/2016/08/22/is-your-smart-security-camera-protecting-your-home-or-spying-on-you/#4fe5341256dd

[3] D. Harwell, "Doorbell-camera firm ring has partnered with 400 police forces, extending surveillance concerns," August 2019. [Online]. Available: https://www.washingtonpost.com/technology/2019/08/28/doorbell-camera-firm-ring-has-partnered-with-police-forces-extending-surveillance-reach/?arc404=true

[4] TechCrunch, "Amazon's camera-equipped echo look raises new questions about smart home privacy," http://alturl.com/7ewnu.

[5] T. U. S. D. of Justice, "Privacy act of 1974, as amended, 5 u.s.c. § 552a," 1974. [Online]. Available: https://www.justice.gov/opcl/privacy-act-1974

[6] M. A. Weiss and K. Archick, "Us-eu data privacy: from safe harbor to privacy shield," 2016.

[7] T. U. S. D. of Homeland Security, "Passenger name records agreements," 2004. [Online]. Available: https://www.dhs.gov/publication/passenger-name-records-agreements

[8] R. Leenes, R. Van Brakel, S. Gutwirth, and P. De Hert, *Data protection and privacy:(In) visibilities and infrastructures*, 2017.

[9] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," in *AAAI*, 2017.

[10] D. J. Butler, J. Huang, F. Roesner, and M. Cakmak, "The privacy-utility tradeoff for remotely teleoperated robots," in *HRI*, 2015.

[11] J. Dai, B. Saghafi, J. Wu, J. Konrad, and P. Ishwar, "Towards privacy-preserving recognition of human activities," in *ICIP*, 2015.

[12] Z. Wu, Z. Wang, Z. Wang, and H. Jin, "Towards privacy-preserving visual recognition via adversarial training: A pilot study," in *ECCV*, 2018.

[13] T. Orekondy, B. Schiele, and M. Fritz, "Towards a visual privacy advisor: Understanding and predicting privacy risks in images," in *ICCV*, 2017.

[14] F. Pittaluga, S. J. Koppal, S. B. Kang, and S. N. Sinha, "Revealing scenes by inverting structure from motion reconstructions," in *CVPR*, 2019.

[15] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks," in *CVPR*, 2016.

[16] H. Kato and T. Harada, "Image reconstruction from bag-of-visual-words," in *CVPR*, 2014.

[17] P. Weinzaepfel, H. Jégou, and P. Pérez, "Reconstructing an image from its local descriptors," in *CVPR*, 2011.

[18] A. Mahendran and A. Vedaldi, "Visualizing deep convolutional neural networks using natural pre-images," *IJCV*, 2016.

[19] C. Gentry *et al.*, "Fully homomorphic encryption using ideal lattices." in *STOC*, 2009.

[20] P. Xie, M. Bilenko, T. Finley, R. Gilad-Bachrach, K. Lauter, and M. Naehrig, "Crypto-nets: Neural networks over encrypted data," *arXiv*, 2014.

[21] A. Chattopadhyay and T. E. Boult, "Privacycam: a privacy preserving camera using uclinux on the blackfin dsp," in *CVPR*, 2007.

[22] T. Winkler, A. Erdélyi, and B. Rinner, "Trusteye. m4: protecting the sensor—not the camera," in *AVSS*, 2014.

[23] Z. W. Wang, V. Vineet, F. Pittaluga, S. N. Sinha, O. Cossairt, and S. Bing Kang, "Privacy-preserving action recognition using coded aperture videos," in *CVPRW*, 2019.

[24] F. Pittaluga and S. J. Koppal, "Privacy preserving optics for miniature vision sensors," in *CVPR*, 2015.

[25] F. Pittaluga and S. J. Koppal, "Pre-capture privacy for small vision sensors," *TPAMI*, 2017.

[26] L. Jia and R. J. Radke, "Using time-of-flight measurements for privacy-preserving tracking in a smart room," *IINF*, 2014.

[27] S. Tao, M. Kudo, and H. Nonaka, "Privacy-preserved behavior analysis and fall detection by an infrared ceiling sensor network," *Sensors*, 2012.

[28] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," *CVPR*, 2016.

[29] B. Cheng, Z. Wang, Z. Zhang, Z. Li, D. Liu, J. Yang, S. Huang, and T. S. Huang, "Robust emotion recognition from low quality and low bit rate video: A deep learning approach," in *ACII*, 2017.

[30] M. Xu, A. Sharghi, X. Chen, and D. J. Crandall, "Fully-coupled two-stream spatiotemporal networks for extremely low resolution action recognition," in *WACV*, 2018.

[31] Z. Wu, K. Suresh, P. Narayanan, H. Xu, H. Kwon, and Z. Wang, "Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach," in *ICCV*, 2019.

[32] P. M Uplavikar, Z. Wu, and Z. Wang, "All-in-one underwater image enhancement using domain-adversarial learning," in *CVPRW*, 2019.

[33] F. Pittaluga, S. Koppal, and A. Chakrabarti, "Learning privacy preserving encodings through adversarial training," in *WACV*, 2019.

[34] M. Bertran, N. Martinez, A. Papadaki, Q. Qiu, M. Rodrigues, G. Reeves, and G. Sapiro, "Adversarially learned representations for information obfuscation and inference," in *ICML*, 2019.

[35] P. C. Roy and V. N. Boddeti, "Mitigating information leakage in image representations: A maximum entropy approach," in *CVPR*, 2019.

[36] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *AIES*, 2018.

[37] Z. Ren, Y. Jae Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," in *ECCV*, 2018.

[38] R. R. Shetty, M. Fritz, and B. Schiele, "Adversarial scene editing: Automatic object removal from weak supervision," in *NeurIPS*, 2018.

[39] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *ICCV*, 2019.

[40] W. Oleszkiewicz, P. Kairouz, K. Piczak, R. Rajagopal, and T. Trzciński, "Siamese generative adversarial privatizer for biometric data," in *ACCV*, 2018.

[41] Y. Li, N. Vishwamitra, B. P. Knijnenburg, H. Hu, and K. Caine, "Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images," in *CVPRW*, 2017.

[42] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele, "Faceless person recognition: Privacy implications in social media," in *ECCV*, 2016.

[43] R. McPherson, R. Shokri, and V. Shmatikov, "Defeating image obfuscation with deep learning," *arXiv*, 2016.

[44] S. J. Oh, M. Fritz, and B. Schiele, "Adversarial image perturbation for privacy protection a game theory perspective," in *ICCV*, 2017.

[45] D. Gurari, Q. Li, C. Lin, Y. Zhao, A. Guo, A. Stangl, and J. P. Bigham, "Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people," in *CVPR*, 2019.

[46] J. Hamm and Y.-K. Noh, "K-beam minimax: Efficient optimization for deep adversarial learning," in *ICML*, 2018.

[47] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015.

[48] X. Xiang and T. D. Tran, "Linear disentangled representation learning for facial actions," *TCSVT*, 2018.

[49] G. Desjardins, A. Courville, and Y. Bengio, "Disentangling factors of variation via generative entangling," *arXiv*, 2012.

[50] A. Gonzalez-Garcia, J. van de Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," *arXiv*, 2018.

[51] S. Reddy, I. Labutov, S. Banerjee, and T. Joachims, "Unbounded human learning: Optimal scheduling for spaced repetition," in *SIGKDD*, 2016.

[52] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

[53] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, 2016.

[55] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv*, 2017.

[56] D.-Z. Du and P. M. Pardalos, *Minimax and applications*, 2013.

[57] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv*, 2012.

[58] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*, 2011.

[59] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *CVPR*, 2018.

[60] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *CVPRW*, 2012.

[61] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *ECCV*, 2016.

[63] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[64] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

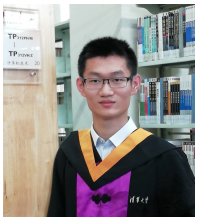
[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[66] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015.

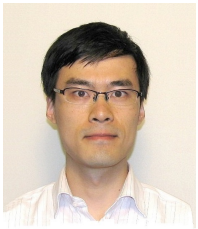
[67] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *CVPR*, 2018.

[68] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv*, 2017.

Zhenyu Wu received the M.S. and B.E. degrees from the Ohio State University and Shanghai Jiao Tong University, in 2017 and 2015 respectively. He is currently a Ph.D. student at Texas A&M University, advised by Prof. Zhangyang Wang. His research interests include privacy/fairness in machine learning, efficient vision, object detection, and adversarial learning.

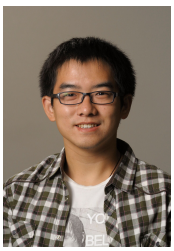


Haotao Wang received the B.E. degree in EE from Tsinghua University, China, in 2018. He is working toward a Ph.D. degree at the University of Texas at Austin, under the supervision of Prof. Zhangyang Wang. His research interests include computer vision and machine learning, especially in fairness/privacy in machine learning, adversarial robustness, and model compression.



Zhaowen Wang received the B.E. and M.S. degrees from Shanghai Jiao Tong University, China, in 2006 and 2009 respectively, and the Ph.D. degree in ECE from UIUC in 2014. He is currently a Senior Research Scientist with the Creative Intelligence Lab, Adobe Inc. His research focuses on understanding and enhancing images, videos and graphics via machine learning algorithms, with a particular interest in sparse coding and deep learning.

Hailin Jin is a Senior Principal Scientist at Adobe Research. He received his M.S. and Ph.D. in EE from WUSTL in 2000 and 2003, respectively. Between fall 2003 and fall 2004, he was a postdoctoral researcher at the CS Department, UCLA. His current research interests include deep learning, computer vision, and natural language processing. His work can be found in many Adobe products, including Photoshop, After Effects, Premiere Pro, and Photoshop Lightroom.



Zhangyang Wang is currently an Assistant Professor of ECE at UT Austin. He was an Assistant Professor of CSE, at Texas A&M University, from 2017 to 2020. He received his Ph.D. degree in ECE from UIUC in 2016, and his bachelor's degree in EEIS from USTC in 2012. Prof. Wang is broadly interested in the fields of machine learning, computer vision, optimization, and their interdisciplinary applications. His latest interests focus on automated machine learning (AutoML), learning-based optimization, machine learning robustness, and efficient deep learning.