Towards Privacy-Preserving Visual Recognition via Adversarial Training: A Pilot Study

Zhenyu Wu[1], Zhangyang Wang[1], Zhaowen Wang[2], and Hailin Jin[2]

[1] Texas A&M University, College Station TX 77843, USA
{wuzhenyu_sjtu,atlaswang}@tamu.edu
[2] Adobe Research, San Jose CA 95110, USA
{zhawang,hljin}@adobe.com

Abstract. This paper aims to improve privacy-preserving visual recognition, an increasingly demanded feature in smart camera applications, by formulating a unique adversarial training framework. The proposed framework explicitly learns a degradation transform for the original video inputs, in order to optimize the trade-off between target task performance and the associated privacy budgets on the degraded video. A notable challenge is that the privacy budget, often defined and measured in task-driven contexts, cannot be reliably indicated using any single model performance, because a strong protection of privacy has to sustain against any possible model that tries to hack privacy information. Such an uncommon situation has motivated us to propose two strategies, i.e., budget model **restarting** and **ensemble**, to enhance the generalization of the learned degradation on protecting privacy against unseen hacker models. Novel training strategies, evaluation protocols, and result visualization methods have been designed accordingly. Two experiments on privacy-preserving action recognition, with privacy budgets defined in various ways, manifest the compelling effectiveness of the proposed framework in simultaneously maintaining high target task (action recognition) performance while suppressing the privacy breach risk. The code is available at https://github.com/wuzhenyusjtu/Privacy-AdversarialLearning

Keywords: Visual privacy, adversarial training, action recognition

1 Introduction

Smart surveillance or smart home cameras, such as Amazon Echo and Nest Cam, are now found in millions of locations to remotely link users to their homes or offices, providing monitoring services to enhance security and/or notify environment changes, as well as lifelogging and intelligent services. Such a prevalence of smart cameras has reinvigorated the privacy debate, since most of them require to upload device-captured visual data to the centralized cloud for analytics. This paper seeks to explore: how to make sure that those smart computer vision devices are only seeing the things that we want them to see (and how to define what we want)? Is it at all possible to alleviate the privacy concerns, without compromising on user convenience?

At the first glance, the question itself is posed as a dilemma: we would like a camera system to recognize important events and assist human daily life by understanding its videos, while preventing it from obtaining sensitive visual information (such as faces) that can intrude people's privacy. Classical cryptographic solutions secure the communication against unauthorized access from attackers. However, they are not immediately applicable to preventing authorized agents (such as the backend analytics) from the unauthorized abuse of information, that causes privacy breach concerns. The popular concept of differential privacy has been introduced to prevent an adversary from gaining additional knowledge by inclusion/exclusion of a subject, but not from gaining knowledge from released data itself [8]. In other words, an adversary can still accurately infer sensitive attributes from any sanitized sample available, which does not violate any of the (proven) properties of differential privacy [18]. It thus becomes a new and appealing problem, to find an appropriate transform on the collected raw visual data at the local camera end, so that the transformed data itself will only enable certain target tasks while obstructing other undesired privacy-related tasks. Recently, some new video acquisition approaches [3,9,47] proposed to intentionally capture or process videos in extremely low-resolution to create privacy-preserving "anonymized videos", and showed promising empirical results.

In contrast, we formulate the privacy-preserving visual recognition in a unique adversarial training framework. The framework explicitly optimizes the trade-off between target task performance and associated privacy budgets, by learning active degradations to transform the video inputs. We investigate a novel way to model privacy budget in a task-driven context. Different from the standard adversarial training where two individual models compete, the privacy budget in our framework cannot be simply defined with one single model, as the ideal protection of privacy has to be universal and model-agnostic, i.e., obstructing every possible model from predicting privacy information. To resolve the so-called **"$\forall$ challenge"**, we propose two strategies, i.e., restarting and ensembling budget model(s), to enhance the generalization capability of the learned degradation to defend against unseen models. Novel training strategies and evaluation protocols have been proposed accordingly. Two experiments on privacy-preserving action recognition, with privacy budgets defined in different ways, manifest the effectiveness of the proposed framework. With many problems left open and large improvement room existing, we hope this pilot study to attract more interests from the community.

2 Related Work

2.1 Privacy Protection in Computer Vision

With pervasive camera for surveillance or smart home devices, privacy-preserving visual recognition has draw increasing interests from both industry and academia, since (1) due to their computationally demanding nature, it is often impractical to run visual recognition tasks all at the resource-limited local device end.

Communicating (part of) data to the cloud is indispensable; (2) while traditional privacy concerns mostly arise from the unsecured channel between cloud and device (e.g, malicious third-party eavesdropping), customers now possess increasing concerns against sharing their private visual information to the cloud (which might turn malicious itself).

A few cryptographic solutions [13,66] were developed to locally encrypt visual information in a homomorphic way, i.e., the cryptosystems allow for basic arithmetic classifiers over encrypted data. However, many encryptions-based solution will incur high computational costs at the local platforms. It is also challenging to generalize the cryptosystems to more complicated classifiers. [4] combined the detection of regions of interest and the real encryption techniques to improve privacy while allowing general surveillance to continue. A seemingly reasonable, and computationally cheaper option is to extract and transmit feature descriptors from raw images, and transmit those features only. Unfortunately, a previous study [31] revealed that considerable information of original images could still be recovered from standard HOG or SIFT features (even they look visually distinct from natural images), making them fragile to privacy hacking too.

An alternative toward a privacy-preserving vision system concerns the concept of anonymized videos. Such videos are intentionally captured or processed to be in special low quality conditions, that only allow for the recognition of some target events or activities, while avoiding the unwanted leak of the identity information for the human subjects in the video [3,9,47]. Typical examples of anonymized videos are videos made to have extreme low resolution (e.g., 16×12) by using low resolution camera hardware [9], based on image operations like blurring and superpixel clustering [3], or introducing cartoon-like effects with a customized version of mean shift filtering [63]. [41,42] 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 [23] explored privacy-preserving tracking and coarse pose estimation using a network of ceiling-mounted time-of-flight low-resolution sensors. [58] adopted a network of ceiling-mounted binary passive infrared sensors. However, both works handled only a limited set of activities performed at specific constrained areas in the room. Later, [47] showed that even at the extreme 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. The usage of low-resolution anonymized videos [9,47] is computationally cheaper, and is also compatible with sensor and bandwidth constraints. However, [9,47] remain empirical in protecting privacy. In particular, neither were their models learned towards protecting any visual privacy, nor were the privacy-preserving effects carefully analyzed and evaluated. In other words, privacy protection in [9,47] came as a "side product" of down-sampling, and was not a result of any optimization. The authors of [9,47] also did not extend their efforts to studying deep learning-based recognition, making their task performance less competitive.

Very recently, a few learning-based approaches have come into play to ensure better privacy protection. [53] defined a utility metric and a privacy metric for a task entity, and then designed a data sanitization function to achieve privacy while providing utility. However, they considered only simple sanitization functions such as linear projection and maximum mean discrepancy transformation. In [43], the authors proposed a game-theoretic framework between an obfuscator and an attacker, in order to hide visual secrets in the camera feed without significantly affecting the functionality of the target application. This seems to be the most relevant work to the proposed one: however, [43] only discussed a toy task to hide QR codes while preserving the overall structure of the image. Another relevant work [18] addressed the optimal utility-privacy tradeoff by formulating it as a min-diff-max optimization problem. Nonetheless, The empirical quantification of privacy budgets in existing works [53,43,18] only considered to protect privacy against *one hacker model*, and was thus insufficient, for which we will explain more in Section 3.1.

2.2 Privacy Protection in Social Media and Photo Sharing

User privacy protection is also a topic of extensive interests in the social media field, especially for photo sharing. The most common means to protect user privacy in a uploaded photo is to add empirical obfuscations, such as blurring, mosaicing or cropping out certain regions (usually faces) [26]. However, extensive research showed that such an empirical means can be easily hacked too [37,32]. A latest work [38] described a game-theoretical system in which the photo owner and the recognition model strive for antagonistic goals of dis-/enabling recognition, and better obfuscation ways could be learned from their competition. However, it was only designed to confuse one specific recognition model, via finding its "adversarial perturbations" [36]. That can caused obvious overfitting as simply changing to another recognition model will likely put the learning efforts in vain: such perturbations even cannot protect privacy from *human eyes*. Their problem setting thus deviated far away from our target problem. Another notable difference is that in social photo sharing, we usually hope to cause minimum perceptual quality loss to those photos, after applying any privacy-preserving transform to them. The same concern does not exist in our scenario, allowing us to explore much more free, even aggressive image distortions.

A useful resource to us was found in [39], which defined concrete privacy attributes and correlated them to image content. The authors categorized possible private information in images, and then run a user study to understand the privacy preferences. They then provided a sizable set of 22k images annotated with 68 privacy attributes, on which they trained privacy attribute predictors.

2.3 Recognition from Visually Degraded Data

To enable the usage of anonymized videos, one important challenge is to ensure reliable performance of the target tasks on those lower-quality videos, besides suppressing the undesired privacy leak. Among all low visual quality scenarios,

visual recognition in low resolution is probably best studied. [61,28,7] showed that low resolution object recognition could be significantly enhanced through proper pre-training and domain adaption. Low-resolution action recognition has also drawn growing interests: [46] proposed a two-stream multi-Siamese CNN that learns the embedding space to be shared by low resolution videos down sampled in different ways, on top of which a transform-robust action classifier was trained. [6] leveraged a semi-coupled filter-sharing two stream network to learn a mapping between the low- and high-resolution feature space. In comparison, the "low-quality" anonymized videos in our case are generated by learned and more complicated degradations, other than simple downsampling [61,6].

3 Technical Approach

3.1 Problem Definition

Assume our training data X (raw visual data captured by camera) are associated with a target task $\mathcal{T}$ and a privacy budget $\mathcal{B}$. We mathematically express the goal of privacy-preserving visual recognition as below (γ is a weight coefficient):

$$\min_{f_T, f_d} L_T(f_T(f_d(X)), Y_T) + \gamma L_B(f_d(X)), \tag{1}$$

where f_T denotes the model to perform the target task $\mathcal{T}$ on its input data. Since $\mathcal{T}$ is usually a supervised task, e.g., action recognition or visual tracking, a label set Y_T is provided on X, and a standard cost function L_T (e.g., softmax) is defined to evaluate the task performance on $\mathcal{T}$. On the other hand, we need to define a budget cost function L_B to evaluate the privacy leak risk of its input data: the larger L_B, the higher privacy leak risk. Our goal is to seek such an *active degradation* function f_d to transform the original X as the common input for both L_T and L_B, such that:

- The target task performance L_T is minimally affected compare to when using the raw data, i.e., $\min_{f_T, f_d} L_T(f_T(f_d(X)), Y_T) \approx \min_{f'_T} L_T(f'_T(X), Y_T)$.
- The privacy budget L_B is greatly suppressed compared to raw data, i.e., $L_B(f_d(X)) \ll L_B(X)$.

The definition of the privacy budget cost L_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 a disclosure of the face or identity of persons. Therefore, to reduce L_B could be interpreted as to suppress the success rate of identity recognition or verification on the transformed video $f_d(X)$. 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 Y_B, and rewrite $L_B(f_d(X))$ as $L_B(f_b(f_d(X)), Y_B)$, where f_b denotes the budget model to predict the corresponding privacy information. *Different from* L_T, minimizing L_B will encourage $f_b(f_d(X))$ to diverge from Y_B as much as possible.

Such a *supervised, task-driven* definition of L_B poses at least two-fold challenges: (1) the privacy budget-related annotations, denoted as Y_B, often have less availability than target task labels. Specifically, it is often challenging to have both Y_T and Y_B ready on the same X; (2) considering the nature of privacy protection, it is not sufficient to merely suppress the success rate of one f_b model. Instead, define a privacy prediction function family $\mathcal{P}$: $f_d(X) \rightarrow Y_B$, the ideal privacy protection of f_d should be reflected as **suppressing every possible model** f_b from $\mathcal{P}$. *That diverts from the common supervised training goal*, where one only needs to find one model to successfully fulfill the target task. We re-write the general form (1) with the task-driven definition of L_B:

$$\min_{f_T, f_d} L_T(f_T(f_d(X), Y_T) + \gamma \max_{f_b \in \mathcal{P}} L_B(f_b(f_d(X)), Y_B). \tag{2}$$

For the solved f_d, the two goals should be simultaneously satisfied: (1) there **exists** ("$\exists$") at least one f_T function that can predict Y_T from $f_d(X)$ well; (2) **for all** ("$\forall$") f_b functions $\in \mathcal{P}$, none of them (even the best one) can reliably predict Y_B from $f_d(X)$. Most existing works chose an empirical f_d (e.g., simple downsampling) and solved $\min_{f_T} L_T(f_T(f_d(X), Y_T)$ [9,61]. [47] essentially solved $\min_{f_T, f_d} L_T(f_T(f_d(X), Y_T)$ to jointly adapted f_d and f_T, after which the authors empirically verified the effect of f_d on L_B (defined as face recognition error rates). Those approaches lack the explicit optimization towards privacy budgets, and thus have no guaranteed privacy-protection effects.

Comparison to Standard Adversarial Training The most notable difference between (2) and existing works based on standard adversarial training [43,38] lies in whether the adversarial perturbations are optimized for "fooling" *one specific* f_b, or *all possible* f_bs. We believe the latter to be necessary, as it considers generalization ability to suppressing unseen privacy breach. Moreover, most existing works seek perturbations with minimal human visual impacts, e.g, by enforcing ℓ_p norm constraint on the pixel domain. That is clearly unaligned with our purpose. In fact, our model could be viewed as to minimize the perturbation in the (learned) feature domain of target utility task.

3.2 Basic Framework

Overview Figure 1 depicts a model architecture to implement the proposed formulation (2). It first takes the original video data X as the input, and passes it through the active degradation module f_d to generate the anonymized video $f_d(X)$. During training, the anonymized video simultaneously goes through a target task model f_T and a privacy prediction model f_b. All three modules, f_d, f_T and f_b, are learnable and can be implemented by neural networks. The entire model is trained under the hybrid loss of L_T and L_B. By tuning the entire pipeline from end to end, $f_d(X)$ will find the optimal task-specific transformation, to the advantage of target task but to the disadvantage of privacy breach, fulfilling the goal of privacy-preserving visual recognition. After training, we can apply the learned active degradation at the local device (e.g., camera) to convert incoming video to its anonymized version, which is then transmitted to the backend (e.g., cloud) for target task analysis.

The proposed framework leads to an adaptive and end-to-end manageable pipeline for privacy-preserving visual recognition. Its methodology is related to the emerging research of feature disentanglement [64]. That technique leads to non-overlapped groups of factorized latent representations, each of which would properly describe information corresponding to particular attributes of interest. Previously it was applied to generative models [10,51] and reinforcement learning [20].

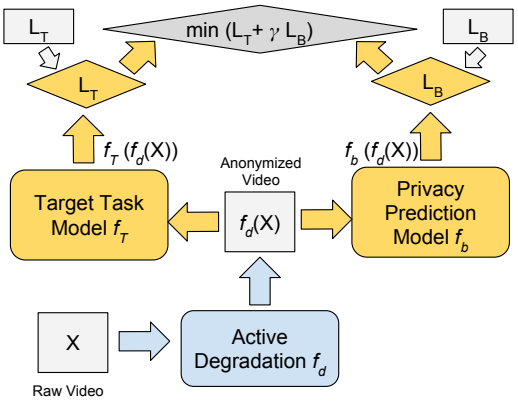


Fig. 1: Basic adversarial training framework for privacy-preserving visual recognition.

Similar to GANs [16] and other adversarial models, our training is prone to collapse and/or bad local minimums. We thus propose a carefully-designed training algorithm with three-module alternating update strategy, explained **in the supplementary**, which could be interpreted as a three-party game. In principle, we strive to avoid any of the three module f_d, f_T, and f_b to change "too quickly", and thus keep monitoring L_T and L_b to decide which of the three modules to be updated next.

Choices of f_d, f_T and f_b The choices of the three modules will significantly impact the performance. As [47] pointed out, f_d can be constructed as a nonlinear mapping by filtering. The form of f_d can be flexible, and its output $f_d(X)$ is unnecessary to be a natural image. For simplicity, we choose f_d to be a "learnable filtering" in the form of 2-D convolutional neural network (CNN), whose the output $f_d(X)$ will be a 2-D feature map of the same resolution as the input video frame. Such a choice is only to facilitate the initial concatenation of building blocks, e.g., f_T and f_b often start with pre-trained models on natural images. Besides, $f_d(X)$ should preferably be in a compact form and light to transmit, considering it will be sent to the cloud through (limited-bandwidth) channels.

To ensure the effectiveness of f_d, it is necessary to choose sufficiently strong f_T and f_b models and let them compete. We employ state-of-the-art video recognition CNNs for corresponding tasks, and adapt them for the degraded input $f_d(X)$ using the robust pre-training strategy proposed in [61].

Particular attentions should be paid towards the budget cost (second term) defined in (2), which we refer as **"the $\forall$ Challenge"**: if we use f_b with some pre-defined CNN architecture, how could we be sure that it is the "best possible" privacy prediction model? That is to say, even we are able to find a f_d function that manages to fail one f_b model, is it possible that some other $f_b' \in \mathcal{P}$ would still be able to predict Y_B from $f_d(X)$, thus leaking privacy? While it is computationally intractable to exhaustively search over $\mathcal{P}$, a naive empirical solution would be to chose a very strong privacy prediction model, hoping that

a f_d function that can confuse this strong one will be able to fool other possible functions as well. However, the resulting $f_d(X)$ may still overfit the artifacts of one specific f_b and fails to generalize. Section 3.3 will introduce two more advanced and feasible recipes.

Choices of L_T and L_B Without loss of generality, we assume both target task f_T and privacy prediction f_b to be classification models and output class labels. To optimize the target task performance, L_T could be simply chosen as the KL divergence: $KL(f_T(f_d(X), Y_T)$.

Choosing L_B is non-standard and tricky since we require minimizing the privacy budget $L_B(f_b(f_d(X)), Y_B)$ to enlarge the divergence between $f_b(f_d(X))$ and Y_B. One possible choice is the negative KL divergence between the predicted class vector and the ground truth label; but minimizing a concave funcion will cause a ton of numerical instabilities (often explosions). Instead, we use the negative entropy function of the predicted class vector, and minimizing it to encourage "uncertain" predictions. Meanwhile, we will use Y_B to ensure a sufficiently strong f_b at the initialization (see 4.1.2). Furthermore, Y_B will play a critical role in model restarting (see 3.3).

3.3 Addressing the $\forall$ Challenge

To improve the generalization of learned f_d 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, e.g., [68], will be explored in future work.

Budget Model Restarting At certain point of training (e.g., when the privacy budget $L_B(f_b(f_d(X)))$ stops decreasing any further), we replace the current weights in f_b with random weights. Such a random re-starting aims to avoid trivial overfitting between f_b and f_d (i.e., f_d is only specialized at confusing the current f_b), without incurring more parameters. We then start to train the new model f_b to be a strong competitor, w.r.t. the current $f_d(X)$: specifically, we freeze the training of f_d and f_T, and change to minimizing $KL(f_b(f_d(X)), Y_B)$, until the new f_b has been trained from scratch to become a strong privacy prediction model over current $f_d(X)$. We then resume adversarial training by unfreezing f_d and f_T, as well as replacing the loss for f_b back to the negative entropy. It can repeat several times.

Budget Model Ensemble The other strategy proposes to approximate the continuous $\mathcal{P}$ with *a discrete set of M sample functions*. Assuming the budget model ensemble $\{f_b^i\}_{i=1}^M$, we turn to minimizing the following discretized surrogate of (2):

$$\min_{f_T, f_d} L_T(f_T(f_d(X), Y_T) + \gamma \max_{i \in \{1,2,\dots,M\}} L_B(f_b^i(f_d(X))). \tag{3}$$

At each iteration (mini-batch), minimizing (3) will only suppress the model f_b^i with the largest L_B cost, e.g., the "most confident" one about its current privacy prediction. The previous basic framework is a special case of (3) with $M = 1$. The ensemble strategy can easily be combined with re-starting.

3.4 Two-Fold Evaluation Protocol

Apart from training data X, assume we have an evaluation set X^e, accompanied with both target task labels Y_T^e and privacy annotations Y_B^e. Our evaluation is significantly more complicated than classical visual recognition problems. After applying the learned active degradation, we need to examine in two folds: (1) whether the learned target task model maintains satisfactory performance; (2) whether the performance of an arbitrary privacy prediction model will deteriorate. The first one can follow the standard routine: applying the learned f_d and f_T to X^e, and computing the classification accuracy A_T via comparing $f_T(f_d(X^e))$ w.r.t. Y_T^e: the higher the better.

For the second evaluation, it is apparently insufficient if we only observe that the learned f_d and f_b lead to poor classification accuracy on X^e, because of the $\forall$ challenge. In other words, f_d needs to generalize not only in the data space, but also w.r.t. the f_b model space. To empirically verify that f_b prohibits reliable privacy prediction for other possible models, we propose a novel procedure: we first re-sample a different set of N models $\{f_b^j\}_{j=1}^N$ from $\mathcal{P}$; none of them will be overlapped with the M budget models used in training. We then train each of them to predict privacy information, over the degraded training data X by applying the learned f_d, i.e., minimizing $f_b^j(f_d(X))$, $j = 1, ..., N$. Eventually, we apply each trained f_b^j and f_d on X^e and compute the classification accuracy for the j-th model. The highest accuracy achieved among the N models on $f_d(X^e)$, denoted as A_b^N, will be by default used to indicate the privacy protection capability of f_d: the lower the better.

4 Experiments

We present two experiments on *privacy-preserving action recognition*, as proof-of-concepts for our proposed general framework. For the target task, we choose video-based action recognition, because it is a highly demanded feature in many smart homes and smart workplaces. The definition of privacy will vary by contexts, and we will study two settings: (1) avoiding the leak of person identities present in the current work; and (2) avoiding the leak of multiple privacy attributes, e.g., multiple crowdsoursed attributes studied in [39]. We emphasize that the generality of the proposed framework (2) can fit in a vast more variety of target task and privacy information compositions.

4.1 Identity-Preserving Action Recognition on SBU

Problem Setting The SBU Kinect Interaction Dataset [67] is a two-person interaction dataset for video-based action recognition, with 8 types of actions and 13 different actor pairs annotated. We define action recognition as the target task $\mathcal{T}$, and the privacy budget task $\mathcal{B}$ as reducing the correct identification rates of the actor pairs in the same video. We note that the target trade-off is highly

challenging to achieve. As can be seen from the first Table **in the supplementary**, the actor pair recognition task easily achieves over 98% accuracy on the original dataset, and stands robust even when the frames are downsampled 28 times, while the action recognition performance already starts to deteriorate significantly. We compare the following five methods:
– **Method 1** (naive downsampling): using raw RGB frames under different down-sampling rates .
– **Method 2** (proposed w/o re-starting): applying the proposed adversarial training to RGB frames, using budget model ensemble *without* restarting .
– **Method 3** (proposed): applying the proposed adversarial training to RGB frames, using budget model ensemble *with* restarting .
– **Method 4**: detecting and cropping out faces from RGB frames.
– **Method 5**: detecting and cropping out whole actor bodies from RGB frames.
Method 1 follows [47], while Methods 4 and 5 are inspired by [26].

Implementation Details We segment video sequences into groups of 16 frames, and use those frame groups as our default input data X. We use the C3D net [60] as the default action recognition model, i.e. f_T. For the f_b identity recognition model, we choose MobileNet [21] to identify actor pair in each frame, and use average pooling to aggregate the frame-wise predictions. The active degradation module f_d adopts the image transformation network in [24].

We choose $\gamma = 2.0$ to suppress the identity recognition performance on SBU. We first initialize the active degradation module f_d as reconstruction of the input. We next take the pre-trained version of C3D net and concatenate it with f_d, and jointly train them for action recognition on the SBU dataset, to initialize f_T. We then freeze them both, and start initializing f_b (MobileNet) for the actor pair identification task, by adapting it to the output of the currently trained f_d. Experiments show that such initializations provide robust starting points for the follow-up adversarial training. If budget model restarting is adopted, we set to "re-start" MobileNet from random initialization after every 100 iterations. The number of ensemble budget models M varies in $\{1, 2, 4, 6, 8, 10, 12, 14, 16, 18\}$. Different budget models can be obtained via setting different depth-multiplier parameter [21] of MobileNet.

Evaluation Procedure We will follow the procedure described in Section 3.4, for two-fold evaluations on the SBU testing set. For the set of models used towards privacy-protection examination, we sample $N = 10$ popular image classification CNNs, a list of which can be found **in the supplementary**. Among them, 8 models start from ImageNet-pretrained versions, including MobileNet (different from those used in training) [21], ResNet [19] and Inception [55]. To eliminate the possibility that the initialization might prohibit privacy prediction, we also intentionally try another 2 models trained from scratch (random initialization). We did not choose any non-CNN image classification model for two reasons: (1) CNNs have state-of-the-art performance and also strong fitting capability when re-trained; (2) most non-CNN image classification models rely on effective feature descriptors, that are designed for natural images. Since $f_d(X)/f_d(X_e)$ are no longer natural images, the effectiveness of such models is in jeopardy too.

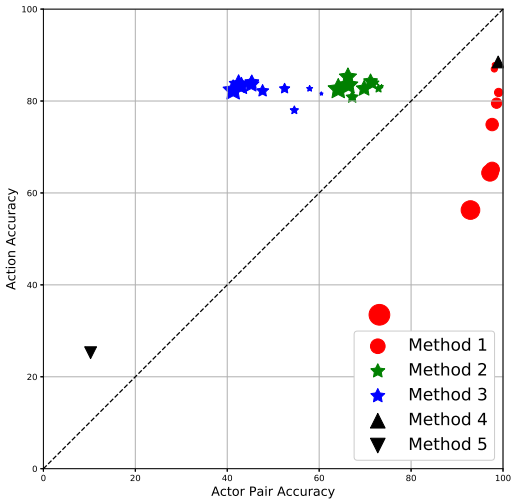


Fig. 2: Target and Budget Task Performance Trade-off on SBU Dataset.

Results and Analysis We present an *innovative visualization* in Figure 2, to display the trade-off between the action recognition accuracy A_T and the actor pair recognition accuracy A_b^N, in an easy-to-interpret way. All accuracy numbers for both task evaluation can be found **in the supplementary**. To read the figure, note that a desirable trade-off should incur minimal loss of A_T (y-axis) when reducing A_b^N (x-axis). Therefore, a point **closer to the upper left corner** denotes a more desirable f_d model that achieves better trade-off, since it incurs less utility performance loss (larger A_T) while suppressing more even the best of N unseen privacy prediction models (smaller A_b^N). For Method 1, a larger marker (circle) size represents a larger downsampling rate. For Methods 2 and 3, a larger marker (star) size denotes more budget models used in ensemble (i.e., larger M). Both Methods 4 and 5 give single points. Observations can be summarized below:

1. Methods 2 and 3 has obvious advantages over naive downsampling (Method 1) in terms of more compelling trade-off between A_T and A_b^N.
2. Method 3 achieves much better individual model performance than Methods 2, showing the significance of **re-starting** in overcoming over-fitting.
3. For Method 3 (similarly for Method 2), adding more f_b models into the budget model ensemble (i.e., increasing M) consistently leads to points closer to the desired upper left (despite small randomness caused by training). The results show that proposed **budget model ensemble** technique effectively alleviates the $\forall$ challenge, making f_d generalize better to unseen f_bs.
4. Neither Methods 4 nor 5 performs competitively. Cropping out faces does not add to protection of identity privacy (its A_b^N is almost the same as using raw RGB frames without downsampling), since body shape and clothing styles reveal significant amounts of identity information too. Cropping out body not only removes identity information, but also deteriorates action recognition.

In brief, Method 3 (the proposed, with both restarting and ensemble) is able to achieve most compelling results on SBU, the best ($M = 18$) at $A_T = 82.3\%$ and $A_b^N = 41.8\%$, without relying on extra hardware (e.g., depth camera). A few visualized examples of learned $f_d(X)$ can be found in Figure 4.

4.2 Protecting Multiple Privacy Attributes on UCF-101/VISPR

Problem Setting The UCF101 dataset [54] is a more challenging action recognition benchmark with 101 different action classes and much richer scenes than

SBU. The newly-released Visual Privacy (VISPR) dataset [39] is a collection of 22, 167 images annotated with 68 privacy attributes, e.g. semi-nudity, hobbies, face, race, gender, skin color, and so on.

Either alone cannot be used for our adversarial training because neither dataset has the pair of competing tasks/labels. However, we show the possibility of jointly utilizing UCF101 and VISPR to train one f_d function, such that it simultaneously enables effective action recognition (as measured on UCF-101) and suppresses multiple privacy attribute prediction (as measured on VISPR), via a novel cross-dataset training and evaluation strategy.

Cross-Dataset Training and Evaluation f_d and f_T are the same as the SBU one. Regarding f_b for the task of multiple privacy attribute prediction, we adopt budget models from {ResNet-V1-50, ResNet-V2-50, MobileNet-1.0, MobileNet-0.75} and vary the number of M in $\{1, 2, 3, 4\}$. They are trained with a multi-label classification loss with sigmoid activations, identically to [39]. γ is chosen as 0.5. Rather than recognition accuracy, the multiple privacy attribute prediction is evaluated by Class-based Mean Average Precision (C-MAP), whose definition can be referred to [39].

The major hassle is the unavailability of $\mathcal{T}$ and $\mathcal{B}$ on the same dataset: since UCF-101 does not have privacy attributes annotated, we cannot directly perform adversarial training and evaluate privacy protection on it; similarly for VISPR. We notice that [39] trained the model on VISPR to be a privacy predictor for general images. We also visually observe that the VISPR model can correctly detect privacy attribute occurring in UCF-101 videos (examples **in the supplementary**). Therefore, we hypothesize that the privacy attributes have good "transferability" between UCF-101 and VISPR, and hence use a privacy prediction model trained on VISPR to assess the privacy leak risk on UCF-101.

Instead of using all 68 attributes in [39], we find that many of them rarely appear in UCF-101 (shown **in the supplementary**). We thus create two subsets for training and evaluating budget models here: one *VISPR-17* set consists of 17 attributes that occurs most in UCF-101 and their associated images in VISPR; the other *VISPR-7* set is further a subset of VISPR-17, that include 7 privacy attributes out of 17 that are most common in smart home settings. Their attribute lists are **in the supplementary**.

During training, we have two pipelines: one is $f_d + f_T$ trained on UCF-101 for action recognition; the other is $f_d + f_b$ trained on VISPR to suppress multiple privacy attribute prediction. The two pipelines *share the same parameters* for f_d. The initialization and alternating training strategy remain unchanged from SBU. During evaluation, we perform the first part of two-fold evaluation, e.g., action recognition, on UCF-101 testing set. We then evaluate the performance of the N-model examination on privacy protection, using the VISPR-17/7 testing sets. Such cross-dataset training and evaluation sheds on new possibilities on training privacy-preserving recognition models, even under the practical shortages of datasets that have been annotated for both tasks.

Results and Analysis We choose Methods 1, 2,and 3 for comparison, defined the same as SBU. All the quantitative results, as well as visualized examples of

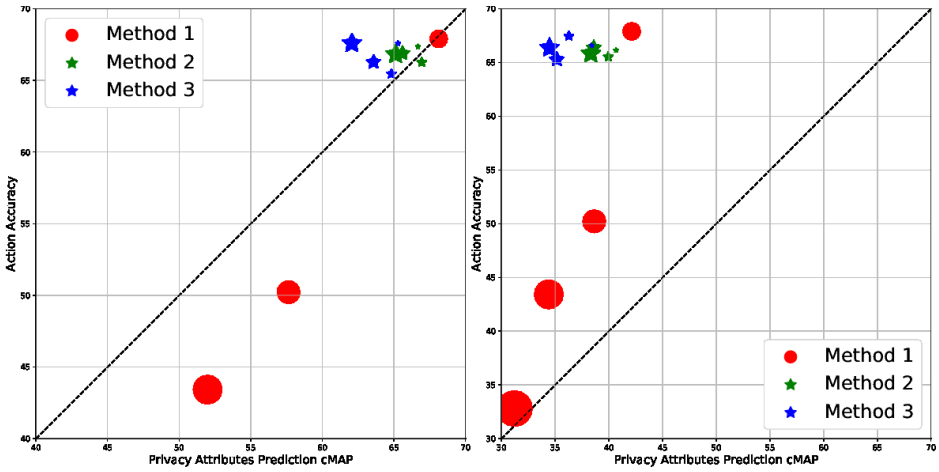


Fig. 3: Performance Trade-off on UCF-101/VISPR dataset. The left one is on VISPR-17 and the right one on VISPR-7.

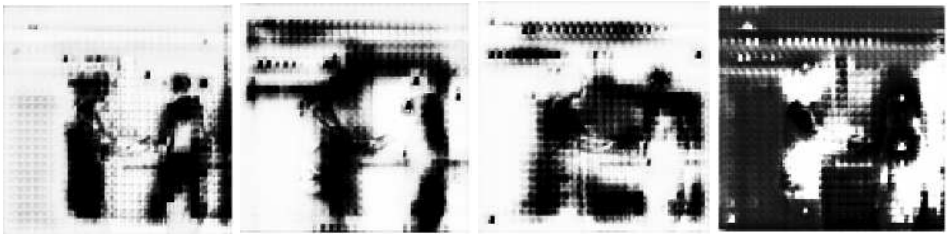
$f_d(X)$ on UCF-101, are shown **in the supplementary**. Similarly to the SBU case, simply downsamping video frames (even with the aid of super resolution as we tried) will not lead to any competitive trade-off between action recognition (at UCF-101) and privacy prediction suppression (at VISPR). As is shown in Figure 3, our proposed adversarial training again leads to more favorable trade-offs on VISPR-17 and VISPR-7, with major conclusions concur with SBU: both ensemble and restarting help f_d generalize better against privacy breach.

5 Limitations and Discussions

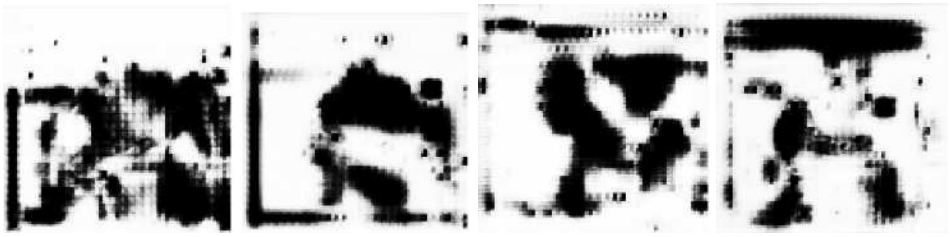
As noted by one anonymous reviewer, a possible alternative to avoid leaking visual privacy to the cloud is to perform action recognition completely at the local device. In comparison, our proposed solution is motivated by at least three folds: **i)** for single utility task (which is not just limited to action recognition), running f_d on device is much more compact and efficient than full f_T For example, our f_T model (11-layer C3D net) has over 70 million parameters, while f_d is a much more compact 3-layer CNN with 1.3 million parameters. At the inference, the total time cost of running f_T over the SBU testing set is 45 times more than running f_d. It also facilitates upgrading to more sophisticated f_T models; **ii)** The smart home scenario calls for the scalability to multiple utility tasks (computer vision functions). It is not economic to load all utility models in the device. Instead, we can train one f_d to work with multiple utility models, and only store and run f_d at the device. More utility models (if no overlap with privacy) could be possibly added in the cloud by training on $f_d(X)$; **iii)** We further point out that the proposed approach can further have a wider practical application scope beyond smart home, e.g, de-identified data sharing.

The current pilot study is preliminary in many ways, and there is large performance room to improve until achieving practical usefulness. First, the definition

Original RGB Frame from UCF-101 (Label: Pushing)



Method 2, M=1 Method 2, M=4 Method 2, M=8 Method 2, M=14



Method 3, M=1 Method 3, M=4 Method 3, M=8 Method 3, M=14

Fig. 4: Example frames after applying the learned degradation on SBU.

of $\mathcal{B}$ and L_B is core to the framework. Considering the $\forall$ challenge, the current budget model ensemble is a rough discretized approximation of $\mathcal{P}$. More elegant ways to tackle this $\forall$ optimization can lead to further breakthroughs in universal privacy protection. Second, adversarial training is well-known to be difficult and instable. Improved training tricks, such as [48], will be useful. Third, a lack of related benchmark datasets, on which $\mathcal{T}$ and $\mathcal{B}$ are both appropriately defined, has become a bottleneck. We see that more concrete and precise privacy definitions, such as VISPR attributes, can certainly result in better feature disentanglement and $\mathcal{T}$-$\mathcal{B}$ performance trade-offs. Current cross-dataset training and evaluation partially alleviate the absence of dedicated datasets. However, the inevitable domain mismatch between two datasets can still hurdle the performance. We plan to refer to crowdsourcing to identify and annotate privacy-related attributes on existing action recognition or other benchmarks, which we hope could help promote this research direction.

References

1. Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.
2. Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*, pages 29–39. Springer, 2011.
3. Daniel J Butler, Justin Huang, Franziska Roesner, and Maya Cakmak. The privacy-utility tradeoff for remotely teleoperated robots. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 27–34. ACM, 2015.
4. Ankur Chattopadhyay and Terrance E Boult. Privacycam: a privacy preserving camera using uclinux on the blackfin dsp. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.
5. Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Action recognition from depth sequences using depth motion maps-based local binary patterns. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 1092–1099. IEEE, 2015.
6. Jiawei Chen, Jonathan Wu, Janusz Konrad, and Prakash Ishwar. Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions. *arXiv preprint arXiv:1610.03898*, 2016.
7. Bowen Cheng, Zhangyang Wang, Zhaobin Zhang, Zhu Li, Ding Liu, Jianchao Yang, Shuai Huang, and Thomas S Huang. Robust emotion recognition from low quality and low bit rate video: A deep learning approach. In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*, pages 65–70. IEEE, 2017.
8. Graham Cormode. Individual privacy vs population privacy: Learning to attack anonymization. *arXiv preprint arXiv:1011.2511*, 2010.
9. Ji Dai, Behrouz Saghafi, Jonathan Wu, Janusz Konrad, and Prakash Ishwar. Towards privacy-preserving recognition of human activities. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 4238–4242. IEEE, 2015.
10. Guillaume Desjardins, Aaron Courville, and Yoshua Bengio. Disentangling factors of variation via generative entangling. *arXiv preprint arXiv:1210.5474*, 2012.
11. Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.
12. Cynthia Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.
13. Zekeriya Erkin, Martin Franz, Jorge Guajardo, Stefan Katzenbeisser, Inald Lagendijk, and Tomas Toft. Privacy-preserving face recognition. In *International Symposium on Privacy Enhancing Technologies Symposium*, 2009.
14. Farhad Farokhi and Henrik Sandberg. Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries. *IEEE Transactions on Smart Grid*, 2017.
15. Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017.

16. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
17. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
18. Jihun Hamm. Minimax filter: learning to preserve privacy from inference attacks. *The Journal of Machine Learning Research*, 18(1):4704–4734, 2017.
19. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
20. Irina Higgins, Arka Pal, Andrei A Rusu, Loic Matthey, Christopher P Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. *arXiv:1707.08475*, 2017.
21. Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
22. Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.
23. Li Jia and Richard J Radke. Using time-of-flight measurements for privacy-preserving tracking in a smart room. *IEEE Transactions on Industrial Informatics*, 10(1):689–696, 2014.
24. Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.
25. Jing Li, Stan Z Li, Quan Pan, and Tao Yang. Illumination and motion-based video enhancement for night surveillance. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 169–175. IEEE, 2005.
26. Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1343–1351. IEEE, 2017.
27. Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.
28. Ding Liu, Bowen Cheng, Zhangyang Wang, Haichao Zhang, and Thomas S Huang. Enhance visual recognition under adverse conditions via deep networks. *arXiv preprint arXiv:1712.07732*, 2017.
29. Ping Liu, Joey Tianyi Zhou, Ivor Wai-Hung Tsang, Zibo Meng, Shizhong Han, and Yan Tong. Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis. In *European Conference on Computer Vision*, pages 151–166. Springer, 2014.
30. Behrooz Mahasseni, Sinisa Todorovic, and Alan Fern. Budget-aware deep semantic video segmentation.
31. Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 2016.
32. Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*, 2016.

33. Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM, 2010.
34. Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on*, pages 173–187. IEEE, 2009.
35. Shree K Nayar and Srinivasa G Narasimhan. Vision in bad weather. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 820–827. IEEE, 1999.
36. Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.
37. Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016.
38. Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial image perturbation for privacy protection–a game theory perspective. In *International Conference on Computer Vision (ICCV)*, 2017.
39. Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.
40. Tribhuvanesh Orekondy, Bernt Schiele, Mario Fritz, and Saarland Informatics Campus. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. *arXiv preprint arXiv:1703.10660*, 2017.
41. Francesco Pittaluga and Sanjeev J Koppal. Privacy preserving optics for miniature vision sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 314–324, 2015.
42. Francesco Pittaluga and Sanjeev Jagannatha Koppal. Pre-capture privacy for small vision sensors. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2215–2226, 2017.
43. Nisarg Raval, Ashwin Machanavajjhala, and Landon P Cox. Protecting visual secrets using adversarial nets. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1329–1332. IEEE, 2017.
44. M. S. Ryoo, T. J. Fuchs, L. Xia, J. K. Aggarwal, and L. Matthies. Robot-centric activity prediction from first-person videos: What will they do to me? In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 295–302, Portland, OR, March 2015.
45. M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, June 2013.
46. Michael S Ryoo, Kiyoon Kim, and Hyun Jong Yang. Extreme low resolution activity recognition with multi-siamese embedding learning. *arXiv preprint arXiv:1708.00999*, 2017.
47. Michael S Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. 2017.
48. Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

49. Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.
50. Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.
51. N Siddharth, Brooks Paige, Alban Desmaison, Jan-Willem van de Meent, Frank Wood, Noah D Goodman, Pushmeet Kohli, and Philip HS Torr. Learning disentangled representations in deep generative models. 2016.
52. Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.
53. Jure Sokolic, Qiang Qiu, Miguel RD Rodrigues, and Guillermo Sapiro. Learning to succeed while teaching to fail: Privacy in closed machine learning systems. *arXiv preprint arXiv:1705.08197*, 2017.
54. Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
55. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
56. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
57. Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
58. Shuai Tao, Mineichi Kudo, and Hidetoshi Nonaka. Privacy-preserved behavior analysis and fall detection by an infrared ceiling sensor network. *Sensors*, 12(12):16920–16936, 2012.
59. TechCrunch. Amazon's camera-equipped echo look raises new questions about smart home privacy. http://alturl.com/7ewnu.
60. Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 4489–4497. IEEE, 2015.
61. Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S Huang. Studying very low resolution recognition using deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
62. Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer vision and image understanding*, 104(2):249–257, 2006.
63. Thomas Winkler, Adám Erdélyi, and Bernhard Rinner. Trusteye. m4: protecting the sensornot the camera. In *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on*, pages 159–164. IEEE, 2014.
64. Xiang Xiang and Trac D Tran. Linear disentangled representation learning for facial actions. *arXiv preprint arXiv:1701.03102*, 2017.
65. Yanchun Xie, Jimin Xiao, Tammam Tillo, Yunchao Wei, and Yao Zhao. 3d video super-resolution using fully convolutional neural networks. In *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, pages 1–6. IEEE, 2016.

66. Ryo Yonetani, Vishnu Naresh Boddeti, Kris M Kitani, and Yoichi Sato. Privacy-preserving visual learning using doubly permuted homomorphic encryption. *arXiv preprint arXiv:1704.02203*, 2017.
67. Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012.
68. Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. *arXiv preprint arXiv:1707.07012*, 2017.