SPAct: Self-supervised Privacy Preservation for Action Recognition

Ishan Rajendrakumar Dave, Chen Chen, Mubarak Shah
Center for Research in Computer Vision, University of Central Florida, Orlando, USA
ishandave@knights.ucf.edu, {chen.chen, shah}@crcv.ucf.edu

Abstract

Visual private information leakage is an emerging key issue for the fast growing applications of video understanding like activity recognition. Existing approaches for mitigating privacy leakage in action recognition require privacy labels along with the action labels from the video dataset. However, annotating frames of video dataset for privacy labels is not feasible. Recent developments of self-supervised learning (SSL) have unleashed the untapped potential of the unlabeled data. For the first time, we present a novel training framework which removes privacy information from input video in a self-supervised manner without requiring privacy labels. Our training framework consists of three main components: anonymization function, self-supervised privacy removal branch, and action recognition branch. We train our framework using a minimax optimization strategy to minimize the action recognition cost function and maximize the privacy cost function through a contrastive self-supervised loss. Employing existing protocols of known-action and privacy attributes, our framework achieves a competitive action-privacy trade-off to the existing state-of-the-art supervised methods. In addition, we introduce a new protocol to evaluate the generalization of learned the anonymization function to novel-action and privacy attributes and show that our self-supervised framework outperforms existing supervised methods. Code available at: https://github.com/DAVEISHAN/SPAct

1. Introduction

Recent advances in action recognition have enabled a wide range of real-world applications, *e.g.* video surveillance camera [7, 24, 35], smart shopping systems like *Amazon Go*, elderly person monitor systems [2, 22, 45]. Most of these video understanding applications involve extensive computation, for which a user needs to share the video data to the cloud computation server. While sharing the videos to the cloud server for the utility action recognition task, the user also ends up sharing the private visual information like gender, skin color, clothing, background objects etc. in the videos as shown in Fig. 1. Therefore, there is a pressing need for solutions to privacy preserving action recognition.

A simple-yet-effective solution for privacy preservation in action recognition is to utilize very low resolution videos (Fig. 1a) [5, 23, 37]. Although this downsampling method does not require any specialized training to remove privacy features, it does not provide a good trade-off between action recognition performance and privacy preservation.

Another set of methods use pretrained object-detectors to detect the privacy regions and then remove or modify the detected regions using synthesis [34] or blurring [47] as shown in Fig. 1b. The detection-based approaches require the bounding-box level annotations for the privacy attributes, and removing the privacy features without an end-to-end learning framework may result in the performance drop of the action recognition task.

Wu *et al*. [41] propose a novel approach to remove the privacy features via learning an *anonymization function* through an adversarial training framework, which requires both *action and privacy labels* from the video. Although the method is able to get a good trade-off of action recognition and privacy preservation, it has two main problems. First, it is not feasible to annotate a video dataset for privacy attributes. For instance, Wu *et al*. [41] acknowledge the issue of privacy annotation time, where it takes immense efforts for them to annotate privacy attributes for even a small-scale (515 videos) video dataset *PA-HMDB*. Second, the learned anonymization function from the known privacy attributes may not generalize in anonymizing the *novel privacy attributes*. For example, in Fig. 1 the learned anonymization function for human-related privacy attributes (*e.g.* gender, skin color, clothing) may still leave other privacy information like scene or background objects un-anonymized.

The performance of the action recognition task depends on the spatio-temporal cues of the input video. Wu *et al*. [41] show that anonymizing the privacy features like face, gender, etc. in the input video does not lead to any reduction in the action recognition performance. Instead of just focusing on the cues based on the privacy annotations, our goal is twofold: 1) learning an anonymization function that can remove all spatial cues in all frames without significantly degrading action recognition performance; and

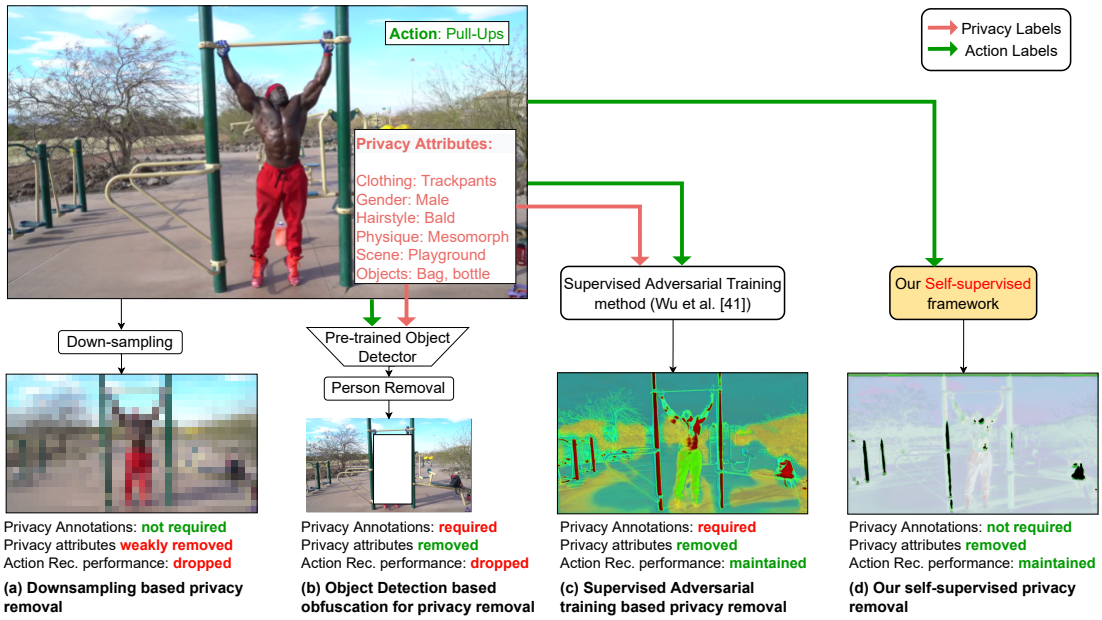

Figure 1. Overview of the existing privacy preserving action recognition approaches. The main goals of a framework include removing privacy information and maintaining action recognition performance at low cost of annotations.

2) learning the anonymization function without any privacy annotations.

Recently, self-supervised learning (SSL) methods have been successfully used to learn the representative features which are suitable for numerous downstream tasks including classification, segmentation, detection, etc. Towards our goal, we propose a novel frame-level SSL method to remove the semantic information from the input video, while maintaining the information that is useful for the action recognition task. We show that our proposed ***S**elf-supervised **P**rivacy-preserving **Act**ion recognition (SPAct)* framework is able to anonymize the video without requiring any privacy annotations in the training.

The learned anonymization function should provide a model-agnostic privacy preservation, hence, we first adopt the protocol from [41] to show the transferability of the anonymization function across different models. *However, there are two aspects in terms of evaluating the generalization ability of the anonymization function, which are overlooked in previous works.*

First, in the real-world scenario, the anonymization function is expected to have *generalization capability with domain shift in action and privacy classes*. To evaluate the generalization capabilities of the anonymization function across novel action and privacy attributes, we propose new protocols. In our experiments, we show that since our model is not dependent on the predefined privacy features like existing supervised methods, and it achieves state-of-the-art generalization across novel privacy attributes.

Second, prior privacy-preserving action recognition works have solely focused on privacy attributes of humans. In practical scenarios, privacy leakage can happen in terms of scene and background objects as well, which could reveal personal identifiable information. Therefore, *the generalization ability of anonymization function to preserve privacy attributes beyond humans* (e.g. *scene and object privacy)* is of paramount importance as well. To evaluate such ability, we propose *P-HVU* dataset, a subset of LSHVU dataset [8], which has multi-label annotations for actions, objects and scenes. Compared to existing same-dataset privacy-action evaluation protocol on PA-HMDB [41], which consists of only 515 test videos, the proposed P-HVU dataset has about 16,000 test videos for robust evaluation of privacy-preserving action recognition.

The contributions of this work are summarized as follows:

- We introduce a novel *self-supervised* learning framework for privacy preserving action recognition without requiring any privacy attribute labels.
- On the existing UCF101-VISPR and PA-HMDB evaluation protocols, our framework achieves a competitive performance compared to the state-of-the-art *supervised* methods which require privacy labels.
- We propose new evaluation protocols for the learned anonymization function to evaluate its generalization capability across novel action and novel privacy attributes. For these protocols, we also show that our method outperforms state-of-the art supervised methods. Finally, we propose a new dataset split *P-HVU* to resolve the issue of smaller evaluation set and extend the privacy evaluation to non-human attributes like action scene and objects.

2. Related Work

Recent approaches for the privacy preservation can be categorized in three major groups: (1) Downsampling based approaches; (2) Obfuscation based approaches; and (3) Adversarial training based approaches. An overview of the existing privacy preserving approaches can be seen in Fig. 1.

Downsampling based approaches utilized a very low resolution input to anonymize the personal identifiable information. Chou *et al*. [4] utilize low resolution depth images to preserve privacy in the hospital environment. Srivastava *et al*. [39] utilize low resolution images to mitigate privacy leakage in human pose estimation. Butler *et al*. [1] use operations like blurring and superpixel clustering to anonymize videos. There are some works [5, 23, 37] utilizing a downsampling based solution for privacy preserving action recognition. An example of anonymization by downsampling is shown in Fig. 1a. Although it is a simple method and does not require privacy-labels for training, one major drawback of the method is its suboptimal trade-off between action recognition and privacy preservation.

Obfuscation based approaches mainly involve using an off-the-shelf object detector to first detect the privacy attributes and then remove or modify the detected regions to make it less informative in terms of privacy features. An interesting solution is proposed by Ren *et al*. [34] for anonymizing faces in the action detection utility. They synthesise a fake image in place of the detected face. A similar approach was taken for the video domain privacy by Zhang *et al*. [47], where first the semantic segmentation is employed to detect the regions of interest, which is followed by a blurring operation to reduce the privacy content of a video. Although the obfuscation based methods work well in preserving the privacy, there are two main problems associated with them: (1) there is domain knowledge required to know the regions of interests, and (2) the performance of the utility task is significantly reduced since this approach is not end-to-end and involves two separate steps: private-object detection/segmentation and object removal.

Recently, Hinojosa *et al*. [17] tackle the privacy preserving human pose estimation problem by optimizing an optical encoder (hardware-level protection) with a software decoder. In addition, some more work focus on *hardware level* protection in the image based vision systems [19, 28, 29, 40], however, they are not within scope of this paper.

Pittaluga *et al*. [27] and Xiao *et al*. [43] propose adversarial optimization strategies for the privacy preservation in the images. Authors in [41, 42] introduce a novel adversarial training framework for privacy preserving action recognition. Their framework adopts a minimax optimization strategy, where action classification cost function is minimized, while privacy classification cost is maximized. Their adversarial framework remarkably outperforms prior works which are based on obfuscation and downsampling.

Recently, self-supervised learning (SSL) based methods have demonstrated learning powerful representations for images [3, 13, 15, 44] and videos [6, 9, 11, 18, 26, 30, 32], which are useful for multiple image and video understanding downstream tasks. In this paper, we propose self-supervised privacy preservation method. Instead of using a privacy classifier to remove only the privacy attributes from the input data like [41], our approach is to remove *all spatial semantic* information from the video, along with keeping the useful *utility action recognition information* by training an anonymization function in an minimax optimization manner. To the bes of our best knowledge, there is no other *self-supervised* privacy preserving action recognition method, which learns in an *end-to-end fashion*, without requiring *privacy labels*.

3. Method

The key idea of our proposed framework is to learn an anonymization function such that it deteriorates the privacy attributes without requiring any privacy labels in the training, and maintains the performance of action recognition task. We build our self-supervised framework upon the existing supervised adversarial training framework of [41]. A schematic of our framework is depicted in Fig. 2. In Sec 3.1, we first formulate the problem by explaining our objective. In Sec 3.2 we present details of each component of our framework, and in Sec 3.3 we explain the optimization algorithm employed in our framework.

3.1. Problem Formulation

Let's consider a video dataset X with action recognition as an utility task, T, and privacy attribute classification as a budget task, B. The goal of a privacy preserving action recognition system is to maintain performance of T, while cutting the budget B. This goal is achieved by learning an anonymization function, f_A, which transforms (anonymize) the original raw data X. Assume that the final system has *any* action classification target model f'_T and *any* privacy target model f'_B. The goal of a privacy preserving training is to find an optimal point of f_A called f^*_A, which is achieved by the following two criteria:

C1: f^*_A should minimally affect the cost function of target model, f'_T, on raw data i.e.

$$L_T(f'_T(f^*_A(X)), Y_T) \approx L_T(f'_T(X), Y_T), \tag{1}$$

where T denotes utility Task, L_T is the loss function which is the standard cross entropy in case of single action label Y_T or binary cross entropy in case of multi-label actions Y_T.

C2: Cost of privacy target model, f'_B, should increase on the transformed (anonymized) data compared to raw data i.e.

$$L_B(f'_B(f^*_A(X))) \gg L_B(f'_B(X)), \tag{2}$$

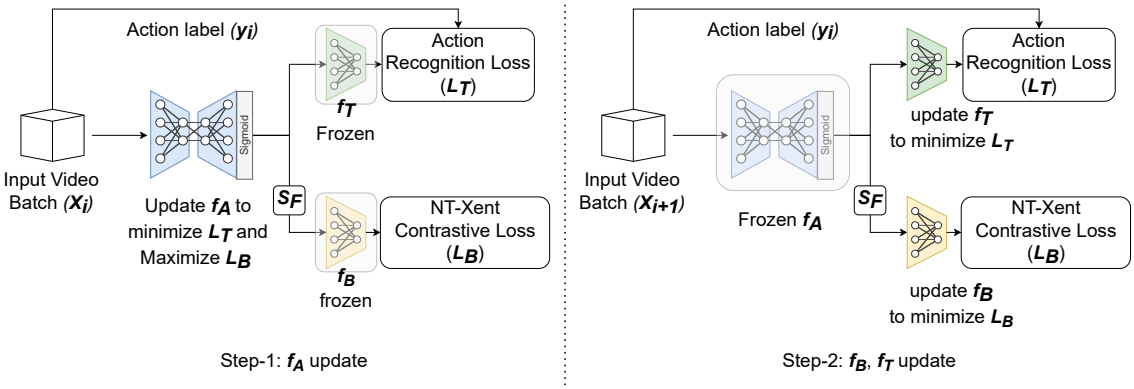

Figure 2. **Minimax optimization in the proposed SPAct framework**. f_A is anonymization function, f_T is 3D-CNN based action classifier, f_B is 2D-CNN based self-supervised learning model, and S_F is temporal sampler. Details of each component can be found in Sec 3.2. We first initialize f_A to identity function and f_T and f_B to pretrained checkpoints optimized on raw video. The proposed minimax optimization strategy is an iterative process including two-steps per iteration. In the left figure, we first update the weights of f_A to minimize action classification loss, L_T, and maximize NT-Xent contrastive self-supervised loss [3] L_B, keeping f_T and f_B frozen. After that as shown in the right figure, for the next batch of videos, we keep f_A frozen and update parameters of f_T and f_B to minimize L_T and L_B, respectively. For more details see Sec 3.3.

where B denotes privacy Budget, L_B is the self-supervised loss for our framework, and binary cross entropy in case of a supervised framework, which requires privacy label annotations Y_B.

Increasing a self-supervised loss L_B results in deteriorating *all* useful information regardless of if it is about privacy attributes or not. However, the useful information for the action recognition is preserved via criterion **C1**. Combining criteria **C1** and **C2**, we can mathematically write the privacy preserving optimization equation as follows, where negative sign before L_B indicates it is optimized by maximizing it:

$$f_A^* = \underset{f_A}{\operatorname{argmin}}[L_T(f_T'(f_A(X)), Y_T) - L_B(f_B'(f_A(X)))]. \tag{3}$$

3.2. Proposed Framework

The proposed framework mainly consists of three components as shown in Fig 2: (1) Anonymization function (f_A); (2) Self-supervised privacy removal branch; and (3) Action recognition or utility branch.

3.2.1 Anonymization Function (f_A)

The anonymization function is a learnable transformation function, which transforms the video in such a way that the transformed information can be useful to learn action classification on *any* target model, f_T', and not useful to learn *any* privacy target model, f_B'. We utilize an encoder-decoder neural network as the anonymization function. f_A is initialized as an identity function by training it using $\mathcal{L}_{L1}$ reconstruction loss as given below:

$$\mathcal{L}_{L1} = \sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} |x_{c,h,w} - \hat{x}_{c,h,w}|, \tag{4}$$

where, x is input image, $\hat{x}$ is sigmoid output of f_A logits, C = input channels, H = input height, and W = input width.

3.2.2 Self-supervised privacy removal branch

A schematic of self-supervised privacy removal branch is shown in Fig. 3. First the video x_i is passed through f_A to get the anonymized video $f_A(x_i)$, which is further passed through a temporal Frame sampler S_F. S_F samples 2 frames out of the video with various S_F strategies, which are studied in Section 5.5. The sampled pair of frames $(S_F(f_A(x_i)))$ are projected into the representation space through 2D-CNN backbone f_B and a non-linear projection head $g(\cdot)$. The pair of frames of video x_i corresponds to projection Z_i and Z_i' in the representation space. The goal of the contrastive loss is to maximize the agreement between projection pair (Z_i, Z_i') of the same video x_i, while maximizing the disagreement between projection pairs of different videos (Z_i, Z_j), where $j \neq i$. The NT-Xent contrastive loss [3] for a batch of N videos is given as follows:

$$L_B^i = -\log \frac{h(Z_i, Z_i')}{\sum_{j=1}^{N}[\mathbb{1}_{[j\neq i]} h(Z_i, Z_j) + h(Z_i, Z_j')]}, \tag{5}$$

where $h(u, v) = \exp\left(u^T v/(\|u\|\|v\|\tau)\right)$ is used to compute the similarity between u and v vectors with an adjustable parameter temperature, τ. $\mathbb{1}_{[j\neq i]} \in \{0, 1\}$ is an indicator function which equals 1 iff $j \neq i$.

3.3. Minimax optimization

In order to optimize the proposed self-supervised framework with the objective of Eq. 3, let's consider anonymization function f_A parameterized by θ_A, and auxiliary models f_B and f_T respectively parameterized by θ_B and θ_T. Assume, α_A, α_B, α_T respectively be the learning rates for θ_A, θ_B, θ_T. First of all, θ_A is initialized as given below (Eq. 6),

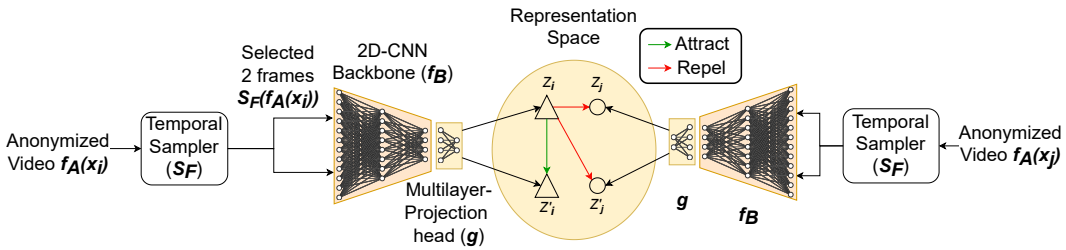

Figure 3. **Contrastive Self-supervised Loss** is used to maximize the agreement between two frames of a video and maximize disagreement between frames of different videos. Please refer to Sec 3.2.2 for more details.

unless f_A reaches to threshold th_{A0} reconstruction performance (Eq. 4) on validation set:

$$\theta_A \leftarrow \theta_A - \alpha_A \nabla_{\theta_A}(\mathcal{L}_{L1}(\theta_A)). \quad (6)$$

Once θ_A is initialized, it is utilized for initialization of θ_T and θ_B as shown in the following equations unless their performance reaches to the loss values of th_{B0} and th_{T0} :

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

$$\theta_B \leftarrow \theta_B - \alpha_B \nabla_{\theta_B}(L_B(\theta_B, \theta_A)). \quad (8)$$

After the initialization, two step iterative optimization process takes place. The first step is depicted in the left side of Fig. 2, where θ_A is updated using the following equation:

$$\theta_A \leftarrow \theta_A - \alpha_A \nabla_{\theta_A}(L_T(\theta_A, \theta_T) - \omega L_B(\theta_A, \theta_B)), \quad (9)$$

where $\omega \in (0, 1)$ is the relative weight of SSL loss, L_B, with respect to supervised action classification loss, L_T. Here the negative sign before L_B indicates that we want to maximize it. In implementation, it can be simply achieved by using negative gradients [10].

In the second step, as shown in the right part of the Fig. 2, θ_T and θ_B are updated using Eq. 7 and 8, respectively. We update θ_B to get powerful negative gradients in the next iteration's step-1. Note that there is a similarity with GAN training here; we can think of f_A as the a generator who tries to fool f_B in the first step and, in the second step f_B tries to get stronger through update of Eq. 8. This two step iterative optimization process continues until L_B reaches to a maximum value th_{Bmax}.

3.4. Intuition: SSL Branch and Privacy removal

Take a model f_b *initialized* with self-supervised contrastive loss (SSL) *pretraining*. Now keeping f_b *frozen*, when we try to *maximize* the contrastive loss, it changes the *input* to f_b in such a way that it decreases *agreement between frames* of the same video. *We know that frames of the same video share a lot of semantic information*, and minimizing the agreement between them results in destroying (i.e. unlearning) most of the semantic info of the input video. In simple terms, maximizing contrastive loss results in destroying all highlighted *attention map* parts of Supp. Fig 7, 8 middle row. Since this unlearned *generic semantic info* contained privacy attributes related to human, scene, and objects; we end up removing *private info* in the input. In this process, we also ensure that semantics related to action reco remains in video, through the action reco branch.

4. Training and Evaluation Protocols

The existing training and evaluation protocols are discussed in Sec 4.1, 4.2 and a new proposed generalization protocol is introduced in Sec 4.3.

4.1. Same-dataset training and evaluation protocol

Training of supervised privacy preserving action recognition method requires a video dataset X^t with action labels Y_T^t, and privacy labels Y_B^t, where t denotes training set. Since, our self-supervised privacy removal framework does not require any privacy labels, we do not utilize Y_B^t. Once the training is finished, the anonymization function is now frozen, called f_A^*, and auxiliary models f_T and f_B are discarded. To evaluate the quality of the learned anonymization, f_A^* is utilized to train: (1) a new action classifier f'_T over the train set $(f_A^*(X^t), Y_T^t)$; and (2) a new privacy classifier f'_B to train over $(f_A^*(X^t), Y_B^t)$. For clarification, we do not utilize privacy labels for training f_A in any protocol. Privacy labels are used only for the evaluation purpose to train the target model f'_B. Once the target models f'_T and f'_B finish training on the anonymized version of train set, they are evaluated on test set $(f_A^*(X^e), Y_T^e)$ and $(f_A^*(X^e), Y_B^e)$, respectively, where e denotes evaluation/test set. Test set performance of the action classifier is denoted as A_T^1 (Top-1 accuracy) or A_T^2 (classwise-mAP), and the performance of privacy classifier is denoted as A_B^1 (classwise-mAP) or A_B^2 (classwise-F_1). Detailed figures explaining different training and evaluation protocols are provided in Supp.Sec.G.

4.2. Cross-dataset training and evaluation protocol

In practice, a trainable-scale video dataset with action and privacy labels doesn't exist. The authors of [41] remedy the supervised training process by a cross-dataset training and/or evaluating protocol. Two different datasets were utilized in [41]: action annotated dataset (X_{action}^t, Y_T^t) to optimize f_A and f_T; and privacy annotated dataset $(X_{privacy}^t, Y_B^t)$ to optimize f_A and f_B. Again, note that in

this protocol, our self-supervised framework does not utilize Y_B^t. After learning the f_A through the different train sets, it is frozen and we call it f_A^*. A new action classifier f_T' is trained on anonymized version of action annotated dataset $(f_A^*(X_{action}^t), Y_T^t)$, and a new privacy classifier f_B' is trained on the anonymized version of the privacy annotated dataset $(f_A^*(X_{privacy}^t), Y_B^t)$. Once the target models f_T' and f_B' finish training on the anonymized version of train sets, they are respectively evaluated on test sets $(f_A^*(X_{action}^e), Y_T^e)$ and $(f_A^*(X_{privacy}^e), Y_B^e)$.

4.3. Novel action and privacy attributes protocol

For the prior two protocols discussed above, the same training set X^t (X_{action}^t and $X_{privacy}^t$) is used for the target models f_T', f_B' and learning the anonymization function f_A. However, a learned anonymization function f_A^* is expected to generalize on any action or privacy attributes. To evaluate the generalization on novel actions, an anonymized verion of novel action set $f_A^*(X_{action}^{nt})$, such that $Y_T^{nt} \cap Y_T^t = \phi$, is used to train the target action model f_T', and its performance is measured on the anonymized test set of novel action set $f_A^*(X_{action}^{ne})$. For privacy generalization evaluation, a novel privacy set $f_A^*(X_{privacy}^{nt})$ (s.t. $Y_B^{nt} \cap Y_B^t = \phi$) (where nt represents novel training) is used to train the privacy target model f_B', and its performance is measured on novel privacy test set $f_A^*(X_{privacy}^{ne})$ (where ne. represents novel evaluation) Please note that novel privacy attribute protocol may not be referred as a *transfer protocol* for the methods, which do not use privacy attributes Y_B^t in learning f_A.

5. Experiments

5.1. Datasets

UCF101 [38] and **HMDB51** [20] are two of the most commonly used datasets for the human action recognition.

PA-HMDB [41] is dataset of 515 videos annotated with video level action annotation and framewise human privacy annotations. The dataset consists of 51 different actions and 5 different human privacy attributes.

P-HVU is a selected subset of LSHVU [8], which is a large-scale dataset of multi-label human actions, with a diverse set of auxiliary annotations provided for object, scenes, concepts, event etc. However, the all auxiliary annotations are not provided for all videos. We select a subset of action-object-scene labels based on their availability in the val set to create our train/test split. The dataset consists of 739 action classes, 1678 objects, and 248 scene categories. Train/test split of P-HVU consists of 245,212/16,012 videos to provide a robust evaluation.

VISPR [25] is an image dataset with a diverse set of personal information in an image like skin color, face, gender, clothing, document information etc.

Further details are provided in Supp.Sec.B.

5.2. Implementation Details

For default experiment setting, we utilize UNet [36] as f_A, R3D-18 [14] as f_T, and ResNet-50 [16] as f_B. For a fair evaluation we report results of different methods with the exact same training augmentations and model architectures. Implementation details related to training setting, hyperparmeters, and model architectures can be found in Supp.Sec.C. Visualization of the learned anonymization from different methods can be seen in Supp.Sec.F.

Downsampling methods are adopted with a down-sampled versions of input resolution with a factor of $2\times$ and $4\times$ used in training and testing.

Obfuscation methods are carried out using a MS-COCO [21] pretrained Yolo [33] object detector to detect person category. The detected persons are removed using two different obfuscation strategies: (1) blackening the detected bounding boxes; (2) applying Gaussian blur in the detected bounding boxes at two different strengths.

5.3. Evaluating learned anonymization on known action and privacy attributes

For known action and privacy attributes, we follow Section 4.2 to evaluate on 2 existing protocols from [41], and follow Section 4.1 to present a new protocol using P-HVU dataset for same-domain training and testing. Results are shown in Table 1.

UCF101-VISPR cross dataset training and evaluation In this protocol, X_{action}^t= UCF101 trainset and X_{action}^e= UCF101 testset; $X_{privacy}^t$ = VISPR trainset and $X_{privacy}^e$= VISPR testset.

HMDB51-VISPR cross dataset training and PA-HMDB evaluation In this protocol, X_{action}^t = HMDB51 trainset, and X_{action}^e = PA-HMDB, $X_{privacy}^t$ = VISPR trainset, and $X_{privacy}^e$ = PA-HMDB.

P-HVU same dataset training and evaluation In this protocol, utility task is multi-label action recognition and privacy is defined in terms of object and scene multi-label classification. In this protocol, X^t = P-HVU trainset, and X^e = P-HVU testset.

We can observe in Table 1 that our proposed *self-supervised* framework achieves a comparable action-privacy trade-off in case of known action and privacy attributes. Other methods like `Downsample-4×`, `Obf-blackening` and `Obf-StrongBlur` get a commendable privacy removal, however, at a cost of action recognition performance.

5.4. Evaluating learned anonymization on Novel action and privacy attributes

Following Sec. 4.3, we propose 2 protocols for the novel actions and 2 protocols for the novel privacy attributes.

Novel action and privacy attributes In this protocol, for actions X_{action}^t = UCF101 trainset, X_{action}^{nt} = HMDB51

Method	UCF101 Action	VISPR1 Privacy		PA-HMDB Action	PA-HMDB Privacy		P-HVU Action	P-HVU Objects	P-HVU Scenes
	Top-1 (↑)	cMAP (↓)	F1 (↓)	Top-1 (↑)	cMAP (↓)	F1 (↓)	cMAP (↑)	cMAP (↓)	cMAP (↓)
Raw data	62.33	64.41	0.555	43.6	70.1	0.401	20.1	11.90	25.8
Downsample-2×	54.11	57.23	0.483	36.1	61.2	0.111	10.9	2.45	8.6
Downsample-4×	39.65	50.07	0.379	25.8	41.4	0.081	0.78	0.89	1.76
Obf-Blackening	53.13	56.39	0.457	34.2	63.8	0.386	8.6	6.12	22.1
Obf-StrongBlur	55.59	55.94	0.456	36.4	64.4	0.243	11.3	6.89	22.8
Obf-WeakBlur	61.52	63.52	0.523	41.7	69.4	0.398	18.6	11.33	25.4
Noise-Features [46]	61.90	62.40	0.531	41.5	69.1	0.384	–	–	–
Supervised [41]	62.10	55.32↓**14%**	0.461↓**17%**	42.3	62.3↓11%	0.194↓51%	18.33	1.98↓83%	9.5↓**63%**
Ours	62.03	57.43↓11%	0.473↓15%	43.1	62.7↓11%	0.176↓**56%**	18.01	1.42↓**88%**	9.91↓62%

Table 1. Comparison of existing privacy preserving action recognition method on **known action and privacy attributes** protocol. Our framework achieves a competitive performance to the supervised method [41]. ↓% denotes relative drop from raw data. For a graphical view, refer to Supp.Sec.D.

Method	Transfer Evaluation: **Action** UCF→HMDB Top-1(%) (↑)	UCF→PA-HMDB Top-1(%) (↑)	Transfer Evaluation: **Privacy** VISPR1→VISPR2 cMAP(%) (↓)	F1 (↓)	Transfer Evaluation P-HVU Action cMAP(%) (↑)	**Scenes → Obj** cMAP(%) (↓)
Raw data	35.6	43.6	57.6	0.498	20.1	11.9
Downsample-2×	24.1	36.1	52.2	0.447	10.9	2.45
Downsample-4×	16.8	25.8	41.5	0.331	0.78	0.89
Obf-Blackening	26.2	34.2	53.6	0.46	8.6	6.12
Obf-StrongBlur	26.4	36.4	53.7	0.462	11.3	6.89
Obf-WeakBlur	33.7	41.7	55.8	0.486	18.6	11.33
Noisy Features [46]	31.2	41.5	53.7	0.458	–	–
Supervised [41]	33.2	40.6	49.6↓14%	0.399↓20%	18.34	6.43↓46%
Ours	34.1	42.8	**47.1**↓**18%**	**0.386**↓**22%**	18.01	**1.42**↓**88%**

Table 2. Comparison of existing privacy preserving action recognition method on **novel action and privacy attributes** protocol. Our framework outperforms the supervised method [41]. ↓% denotes relative drop from raw data.

trainset, X^{ne}_{action} = HMDB51 testset/ PA-HMDB and for privacy, $X^{t}_{privacy}$ = VISPR-1 trainset, $X^{nt}_{privacy}$ = VISPR-2 trainset and $X^{ne}_{privacy}$ = VISPR-2 testset. From the left part of Table 2 and Fig. 4, we can observe that our method outperforms the supervised method [41] in both action and privacy attribute generalization.

Novel privacy attributes from Scenes to Objects In this protocol, we take known action set X^{t}_{action} = P-HVU trainset, and X^{e}_{action} = P-HVU testset, $X^{t}_{privacy}$ = P-HVU trainset Object, $X^{nt}_{privacy}$ = P-HVU trainset Scene and $X^{ne}_{privacy}$ = P-HVU testset Scene. We can observe from the right most part of Table 2 that while testing the learned anononymization from scenes to objects, supervised method [41] gets a similar results like Obf-StrongBlur and removes only ∼**46%** of the raw data's privacy, whereas our method removes ∼**88%** object privacy of the raw data. Main reason for difference in our method's performance gain over [41] in Table 2 is due to the *amount of domain shift* in *novel* privacy attributes. In VISPR1→2, domain shift is very small eg SkinColor(V1)→Tattoo(V2) (Supp.Table 1), and hence [37] is still able to generalize and perform only (<5%) worse than our method. Whereas, in PHVU Scene→Obj, domain shift is huge eg TennisCourt (Scene)→TennisRacket (Obj), where [37] suffers in generalizing and performs significantly (>40%) poor than ours. Additional experiments can be found in Supp.Sec.D and qualitative results can be found in Supp.Sec.F.

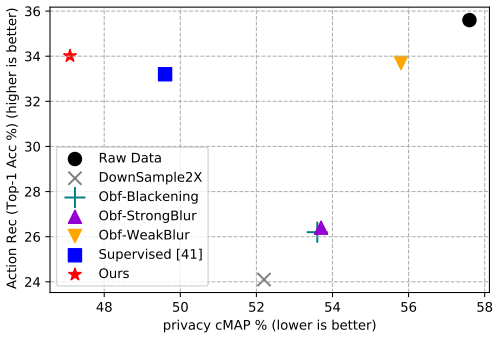

Figure 4. Trade-off between action classification and privacy removal while generalizing from UCF101→HMDB51 for action and VISPR1→VISPR2 for privacy attributes. Our self-supervised method achieves the best trade-off among other methods.

5.5. Ablation Study

Experiments with different privacy removal branches Second row in Table 3 shows the results just using an encoder-decoder based model f_A without using any privacy

removal branch f_B. However, the style changing fails to anonymize privacy information. In our next attempt, we utilize a pretrained SSL *frozen* model to anonymize the privacy information by Eq. 9. This method of frozen f_B is able to remove the privacy information by a small extent ($< 2\%$), however, our biggest boost in privacy removal ($\mathbf{7}\%$) comes from updating f_B with every update in f_A as can be seen in the second last row of Table 3. This observation shows the importance of updating the f_B in step-2 (Eq. 8) of minimax optimization. In other words, we can say that if f_B is not updated with f_A, then it becomes very easy for f_A to fool f_B i.e. maximize L_B, which ultimately leads to a poor privacy removal. Additionally, we also experiment with a spatio-temporal SSL framework as privacy removal branch. Details are given in Supp.Sec.C. Note that removing spatio-temporal semantics from the input video leads to severe degradation in action recognition performance, which is the main reason of choosing 2D SSL privacy removal branch in our framework in order to remove only spatial semantics from the input video.

f_A	f_B	UCF101	VISPR1	
		Top-1 (↑)	cMAP (↓)	F1 (↓)
✗	✗	62.3	64.4	0.555
✓	✗	63.5	64.1	0.549
✓	Spatial (Frozen)	62.2	62.2	0.535
✓	**Spatial**	**62.1**	**57.4**	**0.473**
✓	Spatio-Temporal	56.4	56.6	0.461

Table 3. Experiments with different privacy removal branches

Temporal sampling strategies for SSL In order to experiment with various Temporal sampler (S_F) for choosing a pair of frames from a video, we change the duration (distance) between the two frames as shown in Table 4. The chosen pair of frames from a video is considered for the positive term of contrastive loss (Eq. 5). In our default setting of experiments, we randomly select a pair of frames from a video as shown in the first row. We observe that mining positive frames from further distance decreases the anonymization capability. This is because mining the very dissimilar positives in contrastive loss leads to poorly learned representation, which is also observed while taking temporally distant positive pair in [9, 31].

Distance between positive frames	UCF101 Top-1(%) (↑)	VISPR1 cMAP(%) (↓)	VISPR1 F1 (↓)
No constraint	62.1	57.4	0.473
>64 frames	62.1	58.7	0.488
<8 frames	63.4	57.1	0.443

Table 4. Effect of frame sampling strategy in contrastive loss of SSL privacy removal branch

Effect of different SSL frameworks As shown in Table 5, we experiment with three different 2D SSL schemes in Eq. 5. We can observe that NT-Xent [3] and MoCo [15] achieve comparable performances, however, RotNet [12] framework provides a suboptimal performance in both utility and privacy. Our conjecture is that this is because RotNet mainly encourages learning global representation, and heavily removing the global information from the input via privacy removal branch leads to drop in action recognition performance as well.

SSL Loss	UCF101 Top-1(%) (↑)	VISPR1 cMAP(%) (↓)	VISPR1 F1 (↓)
NT-Xent [3]	62.1	57.4	0.473
MoCo [15]	61.4	57.1	0.462
RotNet [12]	58.1	60.2	0.504

Table 5. Effect of different SSL frameworks

Effect of different f_B and f_T architectures To understand the effect of auxiliary model f_B in the training process of f_A, we experiment with different privacy auxiliary models f_B, and report the performance of their learned f_A^* in the same evaluation setting as shown in Table 6. We can observe that using a better architecture of f_B leads to better anonoymization. There is no significant effect of using different architectures of f_T in learning f_A (Supp.Sec.E).

f_B architecture	UCF101 Top-1 (↑)	VISPR1 cMAP (↓)	VISPR1 F1 (↓)
MobileNetV1 (MV1)	62.1	58.14	0.488
ResNet50 (R50)	62.1	57.43	0.473
R50 + MV1	61.4	56.20	0.454

Table 6. Effect of different f_B in minimax optimization

6. Limitation

One limitation of our work is that it utilizes the basic frameworks for self-supervised learning, and which may be suitable only for the action recognition, and not directly suitable for other video understanding tasks like actions detection or action anticipation. Additionally, there is still room of improvement to match the supervised baseline in case of known action-privacy attributes.

7. Conclusion

We introduced a novel self-supervised privacy preserving action recognition framework which does not require privacy labels for the training. Our extensive experiments show that our framework achieves competitive performance compared to the supervised baseline for the known action-privacy attributes. We also showed that our method achieves better generalization to novel action-privacy attributes compared to the supervised baseline. Our paper underscores the benefits of contrastive self-supervised learning in privacy preserving action recognition.

Acknowledgments

We thank Vishesh Kumar Tanvar, Tushar Sangam, Rohit Gupta, and Zhenyu Wu for constructive suggestions.

References

[1] 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, 2015. 3

[2] Marco Buzzelli, Alessio Albé, and Gianluigi Ciocca. A vision-based system for monitoring elderly people at home. *Applied Sciences*, 10(1):374, 2020. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3, 4, 8

[4] Edward Chou, Matthew Tan, Cherry Zou, Michelle Guo, Albert Haque, Arnold Milstein, and Li Fei-Fei. Privacy-preserving action recognition for smart hospitals using low-resolution depth images. *arXiv preprint arXiv:1811.09950*, 2018. 3

[5] Ji Dai, Behrouz Saghafi, Jonathan Wu, Janusz Konrad, and Prakash Ishwar. Towards privacy-preserving recognition of human activities. In *2015 IEEE international conference on image processing (ICIP)*, pages 4238–4242. IEEE, 2015. 1, 3

[6] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, page 103406, 2022. 3

[7] Ishan Dave, Zacchaeus Scheffer, Akash Kumar, Sarah Shiraz, Yogesh Singh Rawat, and Mubarak Shah. Gabriellav2: Towards better generalization in surveillance videos for action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 122–132, January 2022. 1

[8] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *European Conference on Computer Vision*, pages 593–610. Springer, 2020. 2, 6

[9] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3299–3309, June 2021. 3, 8

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 5

[11] Kirill Gavrilyuk, Mihir Jain, Ilia Karmanov, and Cees GM Snoek. Motion-augmented self-training for video recognition at smaller scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10429–10438, 2021. 3

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 8

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 3

[14] K. Hara, H. Kataoka, and Y. Satoh. Towards good practice for action recognition with spatiotemporal 3d convolutions. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2516–2521, 2018. 6

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[17] Carlos Hinojosa, Juan Carlos Niebles, and Henry Arguello. Learning privacy-preserving optics for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2573–2582, 2021. 3

[18] Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9970–9980, 2021. 3

[19] 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, 2013. 3

[20] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 6

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[22] Jixin Liu, Rong Tan, Guang Han, Ning Sun, and Sam Kwong. Privacy-preserving in-home fall detection using visual shielding sensing and private information-embedding. *IEEE Transactions on Multimedia*, 2020. 1

[23] Jixin Liu and Leilei Zhang. Indoor privacy-preserving action recognition via partially coupled convolutional neural network. In *2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, pages 292–295. IEEE, 2020. 1, 3

[24] Wenhe Liu, Guoliang Kang, Po-Yao Huang, Xiaojun Chang, Yijun Qian, Junwei Liang, Liangke Gui, Jing Wen, and Peng Chen. Argus: Efficient activity detection system for extended video analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, March 2020. 1

[25] 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. 6

[26] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11205–11214, 2021. 3

[27] Francesco Pittaluga, Sanjeev Koppal, and Ayan Chakrabarti. Learning privacy preserving encodings through adversarial training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 791–799. IEEE, 2019. 3

[28] 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. 3

[29] 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, 2016. 3

[30] Rui Qian, Yuxi Li, Huabin Liu, John See, Shuangrui Ding, Xian Liu, Dian Li, and Weiyao Lin. Enhancing self-supervised video representation learning via multi-level feature optimization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[31] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 8

[32] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael Ryoo. Self-supervised video transformer. *arXiv preprint arXiv:2112.01514*, 2021. 3

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 6

[34] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV)*, pages 620–636, 2018. 1, 3

[35] Mamshad Nayeem Rizve, Ugur Demir, Praveen Tirupattur, Aayush Jung Rana, Kevin Duarte, Ishan R Dave, Yogesh S Rawat, and Mubarak Shah. Gabriella: An online system for real-time activity detection in untrimmed security videos. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4237–4244. IEEE, 2021. 1

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6

[37] Michael S Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 1, 3

[38] 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. 6

[39] Vinkle Srivastav, Afshin Gangi, and Nicolas Padoy. Human pose estimation on privacy-preserving low-resolution depth images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 583–591. Springer, 2019. 3

[40] Zihao W Wang, Vibhav Vineet, Francesco Pittaluga, Sudipta N Sinha, Oliver Cossairt, and Sing Bing Kang. Privacy-preserving action recognition using coded aperture videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[41] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 3, 5, 6, 7

[42] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 606–624, 2018. 3

[43] Taihong Xiao, Yi-Hsuan Tsai, Kihyuk Sohn, Manmohan Chandraker, and Ming-Hsuan Yang. Adversarial learning of privacy-preserving and task-oriented representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12434–12441, 2020. 3

[44] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 3

[45] Chenyang Zhang, Yingli Tian, and Elizabeth Capezuti. Privacy preserving automatic fall detection for elderly using rgbd cameras. In *International Conference on Computers for Handicapped Persons*, pages 625–633. Springer, 2012. 1

[46] Dalin Zhang, Lina Yao, Kaixuan Chen, Guodong Long, and Sen Wang. Collective protection: Preventing sensitive inferences via integrative transformation. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1498–1503. IEEE, 2019. 7

[47] Zhixiang Zhang, Thomas Cilloni, Charles Walter, and Charles Fleming. Multi-scale, class-generic, privacy-preserving video. *Electronics*, 10(10):1172, 2021. 1, 3