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
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. Sri- vastava 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 focuses 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 best 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.

$$L_T(f'_T(f'_A(X)), Y_T) \approx L_T(f_T(X), Y_T),$$

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.

$$L_B(f'_B(f'_A(X))) \gg L_B(f'_B(X)),$$
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 $C_1$. Combining criteria $C_1$ and $C_2$, we can mathematically write the privacy preserving optimization equation as follows, where negative sign before $L_B$ indicates it is optimized by maximizing it:

$$f_A^* = \arg\min_{f_A} [L_T(f_T^*(f_A(X)), Y_T) - L_B(f_B(f_A(X)))]$$

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 $L_{L1}$ reconstruction loss as given below:

$$L_{L1} = \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} |x_{c,h,w} - \hat{x}_{c,h,w}|,$$
unles\(f_A\) reaches to threshold \(\theta_{A0}\) reconstruction performance (Eq. 4) on validation set:
\[\theta_A \leftarrow \theta_A - \alpha_A \nabla_{\theta_A} (L_L (\theta_A)).\] (6)
Once \(\theta_A\) is initialized, it is utilized for initialization of \(\theta_T\) and \(\theta_B\) as shown in the following equations unless their performance reaches to the loss values of \(L_{B0}\) and \(L_{T0}\):
\[\theta_T \leftarrow \theta_T - \alpha_T \nabla_{\theta_T} (L_T (\theta_T, \theta_A)),\] (7)
\[\theta_B \leftarrow \theta_B - \alpha_B \nabla_{\theta_B} (L_B (\theta_B, \theta_A)).\] (8)
After the initialization, two step iterative optimization process takes place. The first step is depicted in the left side of Fig. 2, where \(\theta_A\) is updated using the following equation:
\[\theta_A \leftarrow \theta_A - \alpha_A \nabla_{\theta_A} (L_T (\theta_A, \theta_T) - \omega L_B (\theta_B, \theta_A)),\] (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, \(\theta_T\) and \(\theta_B\) are updated using Eq. 7 and 8, respectively. We update \(\theta_B\) 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 \(\theta_{Bmax}\).

### 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.

### 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 \(Y^t\), and privacy labels \(Y^p\), where \(t\) denotes training set. Since, our self-supervised privacy removal framework does not require any privacy labels, we do not utilize \(Y^p\). Once the training is finished, the anonymization function is now frozen, called \(f_A^\prime\), and auxiliary models \(f_T\) and \(f_B\) are discarded. To evaluate the quality of the learned anonymization, \(f_A^\prime\) is utilized to train: (1) a new action classifier \(f_T^1\) over the train set \((f_A^\prime (X^t), Y^t_1)\); and (2) a new privacy classifier \(f_B^1\) to train over \((f_A^\prime (X^t), Y^p_1)\). 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 \(f_B^1\). Once the target models \(f_T^1\) and \(f_B^1\) finish training on the anonymized version of train set, they are evaluated on test set \((f_A^\prime (X^t), Y^t_2)\) and \((f_A^\prime (X^t), Y^p_2)\), respectively, where \(e\) denotes evaluation/test set. Test set performance of the action classifier is denoted as \(A^1_T\) (Top-1 accuracy) or \(A^2_T\) (classwise-mAP), and the performance of privacy classifier is denoted as \(A^1_B\) (classwise-mAP) or \(A^2_B\) (classwise-F1). 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 \((X^t_{\text{action}}, Y^t_1)\) to optimize \(f_A\) and \(f_T\); and privacy annotated dataset \((X^t_{\text{privacy}}, Y^p_1)\) to optimize \(f_B\). Again, note that in
this protocol, our self-supervised framework does not utilize \( Y_B \). After learning the \( f_A \) through the different train sets, it is frozen and we call it \( f_A^\ast \). A new action classifier \( f_T^\ast \) is trained on anonymized version of action annotated dataset \((f_A^\ast(X_{\text{action}}^t), Y_B^t)\), and a new privacy classifier \( f_B^\ast \) is trained on the anonymized version of the privacy annotated dataset \((f_A^\ast(X_{\text{privacy}}^t), Y_B^t)\). Once the target models \( f_T^\ast \) and \( f_B^\ast \) finish training on the anonymized version of train sets, they are respectively evaluated on test sets \((f_A^\ast(X_{\text{action}}^e), Y_B^t)\) and \((f_A^\ast(X_{\text{privacy}}^e), Y_B^t)\).

4.3. Novel action and privacy attributes protocol

For the prior two protocols discussed above, the same training set \( X^t \) (\( X_{\text{action}}^t \) and \( X_{\text{privacy}}^t \)) is used for the target models \( f_T^\ast \), \( f_B^\ast \) and learning the anonymization function \( f_A^\ast \). However, a learned anonymization function \( f_A^\ast \) is expected to generalize on any action or privacy attributes. To evaluate the generalization on novel actions, an anonymized version of novel action set \( f_A^\ast(X_{\text{action}}^{nt}) \), such that \( Y_B^{nt} \cap Y_B^t = \phi \), is used to train the target action model \( f_T^\ast \), and its performance is measured on the anonymized test set of novel action set \( f_A^\ast(X_{\text{action}}^{nt}) \). For privacy generalization evaluation, a novel privacy set \( f_A^\ast(X_{\text{privacy}}^{nt}) \) (s.t. \( Y_B^{nt} \cap Y_B^t = \phi \)) is used to train the privacy target model \( f_B^\ast \), and its performance is measured on novel privacy test set \( f_A^\ast(X_{\text{privacy}}^{nt}) \) (where \( nt \) 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 a 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, hyperparameters, 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, \( X_{\text{action}}^t = \) UCF101 trainset and \( X_{\text{action}}^e = \) UCF101 testset; \( X_{\text{privacy}}^t = \) VISPR trainset and \( X_{\text{privacy}}^e = \) VISPR testset.

HMDB51-VISPR cross dataset training and PA-HMDB evaluation

In this protocol, \( X_{\text{action}}^t = \) HMDB51 trainset, and \( X_{\text{action}}^e = \) PA-HMDB, \( X_{\text{privacy}}^t = \) VISPR trainset, and \( X_{\text{privacy}}^e = \) VISPR testset.

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 Downsampling-4x, 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 \( X_{\text{action}}^t = \) UCF101 trainset, \( X_{\text{action}}^e = \) HMDB51 trainset, and privacy attributes. Other methods like Downsampling-4x, Obf-blackening and Obf-StrongBlur get a commendable privacy removal, however, at a cost of action recognition performance.
In this protocol, we take known action set $X^{act}_\text{train}$ = P-HVU trainset, and $X^{pr}_{\text{privacy}}$ = VISPR-2 trainset. $X^{pr}_{\text{privacy}}$ = VISPR-2 trainset Object, $X^{pr}_{\text{privacy}}$ = P-HVU trainset Scene and $X^{pr}_{\text{privacy}}$ = VISPR-2 trainset Scene. We can observe from the right most part of Table 2 that while testing the learned anonymization from scenes to objects, supervised method [41] gets a similar results like Obf-StrongBlur and removes only $\sim 46\%$ of the raw data’s privacy, whereas our method removes $\sim 88\%$ 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 $\rightarrow$ VISPR2, domain shift is very small eg SkinColor(V1) $\rightarrow$ 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 $\rightarrow$ Obj, domain shift is huge eg TennisCourt (Scene) $\rightarrow$ 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.

### 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 branches, which shows that the best performance is achieved with a combination of all branches.

### 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.

| Method | UCF101 | VISPR1 | PA-HMDB | P-HVU |
|--------|--------|--------|---------|-------|
|        | Action | Privacy | Action | Privacy | Action | Privacy |
|        | Top-1 (%) | cMAP (%) | Top-1 (%) | cMAP (%) | Top-1 (%) | cMAP (%) |
| Raw data | 35.6 | 43.6 | 57.6 | 0.498 | 20.1 | 11.9 |
| Downsample-2x | 24.1 | 36.6 | 52.2 | 0.447 | 10.9 | 2.45 |
| Downsample-4x | 16.8 | 25.8 | 41.5 | 0.331 | 0.78 | 0.89 |
| Obf-Blackening | 26.2 | 34.5 | 53.6 | 0.46 | 8.6 | 6.12 |
| Obf-StrongBlur | 26.4 | 36.5 | 53.7 | 0.466 | 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 | 0.399 | 18.33 | 1.98 |
| Ours | 34.1 | 42.8 | 47.1 | 0.386 | 18.01 | 1.42 |

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

20138
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 (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.

Effect of different SSL frameworks

In order to experiment with various Temporal sampler ($S_T$) 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].

| $f_A$ | $f_B$ | UCF101 Top-1(↑) | VISPRI1 cMAP(↓) | VISPRI1 F1(↓) |
|-------|-------|----------------|----------------|---------------|
| ✓     | ✓     | 62.3          | 64.4           | 0.535         |
| ✓     | X     | 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

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(↑) | VISPRI1 cMAP(↓) | VISPRI1 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 anonymization. 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(↑) | VISPRI1 cMAP(↓) | VISPRI1 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.
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 Jagannatha 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. Precapture privacy for small vision sensors. IEEE transactions on pattern analysis and machine intelligence, 39(11):2215–2226, 2016. 3

[30] Rui Qian, Yu Xi Li, Huabin Liu, John See, Shuangrui Ding, Xian Liu, Dian Li, and Wei Yao 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 620–630, 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] Zhiwo 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 Zhangyong 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, Zhangyong 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, Mannohan 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 Cillonzi, Charles Walter, and Charles Fleming. Multi-scale, class-generic, privacy-preserving video. Electronics, 10(10):1172, 2021. 1, 3