TeD-SPAD: Temporal Distinctiveness for Self-supervised Privacy-preservation for video Anomaly Detection

Anonymous ICCV submission

Paper ID 2874

Abstract

Video anomaly detection (VAD) without human monitoring is a difficult computer vision task that can have a positive impact on society if implemented successfully. While recent advances have made significant progress in solving this task, most existing approaches overlook a critical real-world concern: privacy. With the increasing popularity of artificial intelligence technologies, it becomes crucial to implement proper AI ethics into their development. Privacy leakage in VAD allows models to pick up and amplify unnecessary biases related to people's personal information, which may lead to undesirable decision making. In this paper, we propose TeD-SPAD, a privacy-aware video anomaly detection framework that destroys visual private information in a self-supervised manner. In particular, we explore the impact of temporally-distinctive video representations for VAD, finding that temporal distinctiveness pairs well with current anomaly feature representation learning methods. We achieve a positive trade-off between privacy protection and utility anomaly detection performance on three popular weakly supervised VAD datasets: UCF-Crime, XD-Violence, and ShanghaiTech. Our proposed anonymization model reduces private attribute prediction by 32.25% while only reducing frame-level ROC AUC on the UCF-Crime anomaly detection dataset by 3.69%.

1. Introduction

Machine learning-driven technologies are increasingly being adopted by society. The progress in cloud computing has enabled the deployment of even computationally intensive technologies in the public space. One such application is video anomaly detection (VAD) in autonomous surveillance. VAD is a video understanding task that aims to identify the temporal location of anomalous events occurring in long continuous videos without human supervision. An anomaly can be defined as any unusual event, such as shoplifting, fighting, or vandalism. Proper application of this technology can result in faster response times to anomalies like crimes, without the need for human resources to monitor camera feeds.

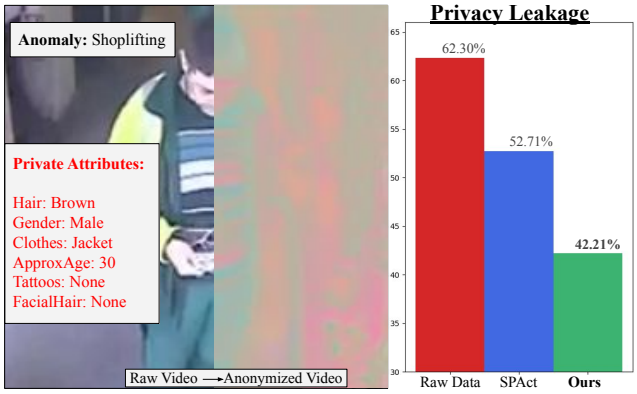


Figure 1: Single frame from video Shoplifting052_x264.mp4 of UCF-Crime [40] showing the types of private attributes leaked in visual data. The graph demonstrates the ability of our self-supervised framework to mitigate privacy leakage. At a similar anomaly-detection utility performance, our method prevents **32%** of the visual private leakage compared to raw data.

However, public adoption of such AI technologies brings justifiable concern about safety and their decision-making capabilities. Many of these concerns center around privacy violations and accuracy. VAD in autonomous surveillance is one such application where visual *privacy leakage* and *data bias* are exceedingly important issues. Sending videos to remote computers or cloud services to process results in unnecessary privacy leakage for people who are not directly involved in anomalous events. Additionally, an application employing a standard RGB video will incorporate any bias found in its training set, leading to potentially unfair decisions. An illustration of privacy leakage is shown in Fig. 1.

Recently, there have been interesting attempts to prevent visual privacy leakage in action recognition. Some of the approaches utilize input downsizing-based solutions [36, 8, 25] and object-detection-dependent obfuscation-based formulations [34, 50]. Recently, Wu *et al*. [46] proposed an adversarial training-based framework where they train

an anonymization function to remove privacy preservation. Dave *et al*. [11] proposed a self-supervised privacy-preserving framework that does not require privacy labels and achieves similar performance to the supervised method [46]. Since many weakly-supervised anomaly detection (WSAD) methods rely on the fixed features of action recognition, privacy-preserving action recognition seems like a promising candidate for privacy-preserving anomaly detection. However, detecting anomalies does not align well with privacy-preserved action recognition videos. This is because WSAD involves localizing anomalies within long and untrimmed videos, which requires more temporal reasoning than the utility task of action recognition.

To the best of our knowledge, privacy preservation in video anomaly detection is an unexplored area in computer vision. Building on the existing self-supervised privacy-preserving action recognition framework [11], we propose a more aligned utility branch for anomaly detection. Detecting anomalies in long, untrimmed videos requires *temporally distinctive* reasoning to determine whether events in the *same scene* are anomalous. This is why most existing anomaly detection methods focus on refining the features of pre-trained video encoders to increase their temporal separability. To achieve this, we use a novel temporal-distinctive triplet loss to promote temporal distinctiveness during anonymization training. Once the anonymization function is learned through our proposed anonymization framework, we apply it to the anomaly dataset, which ensures privacy leakage mitigation in privacy-sensitive anomaly detection tasks. We use these anonymized features to train the state-of-the-art WSAD method, MGFN [6].

To evaluate privacy-preserving performance anomaly detection, we adopt protocols from prior action recognition methods, where we report the utility performance on WSAD task on widely used anomaly datasets (UCF-Crimes [40], XD-Violence [45], and ShanghaiTech [26]) and budget performance on privacy dataset VISPR [29].

Our contributions can be summarized as follows:

- We introduce a new problem of privacy preservation in video anomaly detection, where we identify the privacy leakage issue in existing weakly supervised anomaly detection methods.
- To address this open problem, we devise TeD-SPAD, a framework based on self-supervised privacy preservation with a temporal-distinctiveness objective to make the video anonymization process more suitable for anomaly detection.
- We propose evaluation protocols for privacy vs anomaly trade-off, and demonstrate that our proposed framework outperforms prior methods by significant margins across all anomaly detection benchmarks. On the widely used UCF-Crimes dataset, our method is able to eliminate **32.25%** of the privacy leakage at a cost of only a 3.96% reduction in frame-level AUC performance.

2. Related Works

Privacy Preservation

We observe that many works preserve visual privacy at capture time by using non-intrusive sensors such as thermal imaging, depth cameras, or event cameras [27, 16, 19]. Other works allow for raw RGB visual information to be captured, but make an effort to protect the subject privacy in such a way that the data is still useful in a utility task. Earlier efforts aimed at dealing with visual privacy include image downsampling [9] or blocking/blurring privacy-related objects located using pretrained object detectors. Both of these obfuscation methods were shown to reduce utility results by more than they reduced privacy leakage [11, 22, 46]. Instead of blocking or blurring, [38] creatively uses a generative adversarial network (GAN) to inpaint removed private objects from images. Syfer [47] uses a random encoding scheme that prevents patient identification in medical images while maintaining utility task performance.

Wu *et al*. released an action dataset with privacy labels, PA-HMDB [46]. They use an adversarial learning framework to obfuscate the privacy features using supervised privacy labels. MaSS [2] uses a similar framework to Wu *et al*. [46], except it adapts a compound loss to flexibly preserve certain attributes instead of destroying them. STPrivacy [22] upgraded the general framework to work with a transformer anonymizing block, masking entire video tubelets unnecessary for action recognition. Following Dave *et al*.'s SPAct [11], we adopt a similar self-supervised adversarial anonymization framework without the use of the privacy labels, using a negative NT-Xent [4] contrastive loss to destroy spatial privacy information.

Anomaly Detection

With such a high volume of available video footage, it is infeasible to create sufficient labelled data to solve supervised VAD. Therefore, many works explore unsupervised methods. These generally train a reconstruction model, then either reconstruct the current frame or try to predict the next frame, signaling an anomaly when reconstruction error is high [7, 31, 21, 30, 48, 41]. Giorgi *et al*. [14] used an autoencoder with differential privacy, generating anomaly scores from the noisy reconstructed images. This method helped retain some level of subject privacy, but was only evaluated on image quality metrics.

Sultani *et al*. [40] brought weak supervision to VAD, where anomalies are labelled at the video level, by introducing UCF-Crime, a large scale weakly supervised dataset. The authors propose formulating weakly supervised VAD as a multi instance learning (MIL) problem, showcasing the

benefits of temporal smoothness loss & sparsity loss. With the exception of a select few works [51, 49], all following weakly supervised methods are considered anomaly feature representation learning as they improve upon MIL formulation [52, 42, 13, 23, 17, 43, 6, 37], which involves interpreting static video features extracted using an action classifier. Zhong *et al*. [51] is one such exception, they reformulated the problem as binary classification with noisy labels, adopting a graph convolutional network (GCN) to correct the labels. On top of this, the authors proposed a rearranged weakly supervised version of ShanghaiTech [26]. Wu *et al*. [45] introduce XD-Violence, an even larger weakly supervised dataset that includes audio and focuses on violent anomalous activities, bringing multi-modal fusion to VAD. In this work, we choose to focus on the weakly supervised video anomaly detection setting due to its effectiveness and low annotation effort.

Most works find it useful to model temporal relations between video segments [42, 13, 28, 45, 23, 43, 52, 3]. Since anomalous segment features tend to have larger feature magnitudes than normal segments [42, 6], [28] is able to exploit dynamic variations in features between consecutive segments to help localize anomalies. Complementing this idea, [44, 42, 6] demonstrate the effectiveness of explicitly encouraging feature discrimination. Intuitively, these observations can be aggregated by enforcing temporally distinct feature representations.

3. Method

The central idea behind our proposed framework is to develop an anonymization function that can degrade privacy attributes during training without relying on privacy labels. Furthermore, this function must be able to maintain the performance of the weakly-supervised anomaly detection task. Fig. 2 displays a schematic diagram of the proposed framework. In Sec. 3.1, we provide a comprehensive discussion of the problem statement. Next, in Sec. 3.2, we describe the component of the framework and their initialization process. Sec. 3.3 outlines the anonymization function training, where we propose a temporal-distinctive triplet loss to enhance the existing self-supervised privacy-preserving framework [11]. Once we learn the anonymization function, in Sec. 3.4, we train the anomaly detection model using the privacy-preserved features obtained through our anonymization function. An overview of our complete framework is outlined in Algorithm Sec. 3.5.

3.1. Problem Statement

Suppose we have a video dataset $\mathbb{D}_{anomaly} = \{X^{(i)}, Y^{i}\}_{i=1}^{N}$, where, $X^{(i)}$ is a video-instance, N is a total number of samples and $Y^{(i)} \in \{0, 1\}$ is a binary label. Considering video-level anomaly detection as a utility task T, and privacy attribute classification as the budget task B, the aim of a privacy-preserving system is to maintain the performance of T while reducing B. To achieve this goal, the system learns an anonymization function f_A, which modifies the original raw data. This goal of privacy-preservation could be fundamentally expressed as following criteria:
Criterion-1: The performance of the utility task should not degrade from the original performance, i.e loss $\mathcal{L}_T$ value of the utility target model f'_T should remain almost identical before and after applying the anonymization function.

$$\mathcal{L}_T(f'_T(f^*_A(X)), Y) \approx \mathcal{L}_T(f'_T(X), Y), \tag{1}$$

Criterion-2: Applying anonymization function should increase the loss $\mathcal{L}_B$ for budget B task of target budget model f'_B.

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

Regarding weakly-supervised anomaly detection (WSAD), most of the existing methods require multi-stage training, meaning they are not end-to-end. This presents a challenge for incorporating it as a utility task in anonymization training. By contrast, privacy-preserving action recognition frameworks [11, 46] have an end-to-end utility task (i.e action recognition), making it more straightforward to include.

Since most WSAD methods rely on fixed video encoder features from large-scale action recognition training, we can utilize the exact same anonymization process of privacy-preserving action recognition [11], utilizing action recognition as a *proxy-utility task* on a proxy-utility action dataset $\mathbb{D}_{action}$. However, the problem with such anonymization training is that the proxy-utility task (action recognition) is not well aligned with the true utility task of anomaly detection. Because of that, anomaly detection performance drops significantly as the training progress. To resolve this issue, we reformulate its utility branch with temporal-distinctiveness to better align with anomaly detection task.

3.2. Anonymization Framework

Our anonymization framework consists of 3 main components:(1) Anonymization function (f_A), which is a simple encoder-decoder model with a sigmoid activation (2) Privacy removal model (f_B), which is an image encoder, and (3) Utility model (f_T), which is a video encoder.
Framework Initialization First of all, our anonymization model is pre-trained to initialize as an identity function. This pre-training involves the reconstruction of the frames from $\mathbb{D}_{action}$ using L1-reconstruction loss.

$$\mathcal{L}_{L1} = \sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} |x_{c,h,w} - \hat{x}_{c,h,w}|, \tag{3}$$

where $\hat{x}$ is the output of f_A, x is an input image, and C, H, W corresponds to the channel, height, and width of the input image.

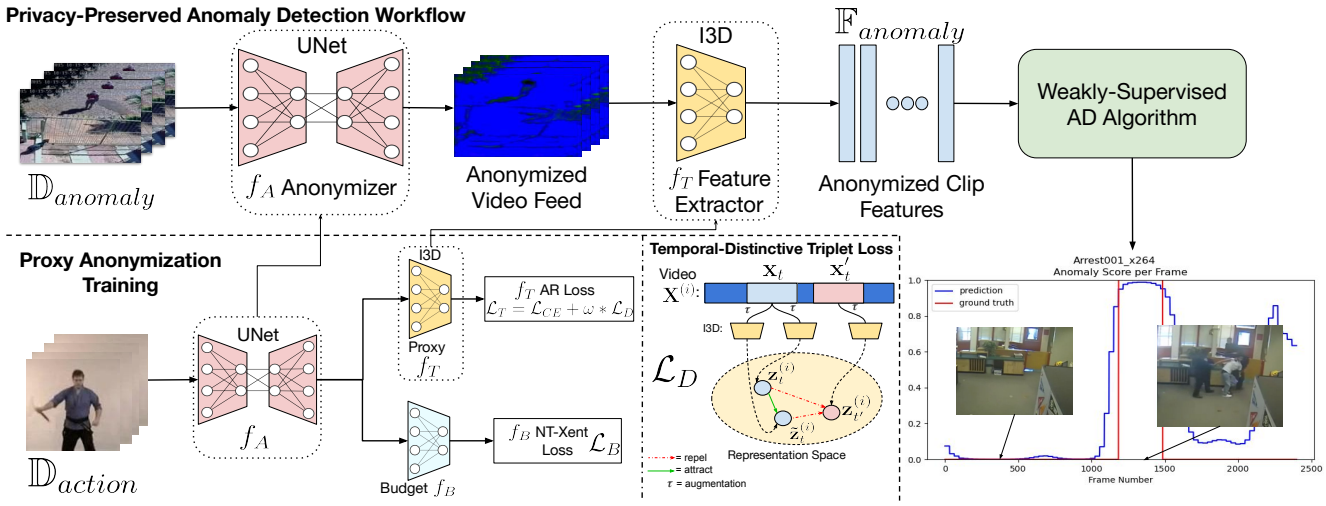


Figure 2: Diagram of the privacy-preserved anomaly detection workflow along with the proxy anonymization training. The lower half shows training the UNet anonymizer and the I3D action classifier along with the temporal-distinctive triplet loss. After training the anonymizer and feature extractor, the now anonymized anomaly dataset videos are passed through any WSAD algorithm, outputting frame level anomlay scores for evaluation.

Secondly, the privacy model f_B is initialized with the self-supervised weights of SimCLR [5] on ImageNet [12]. And, video encoder f_T is pre-trained with the standard action recognition weights from Kinetics400 dataset [1].

3.3. Anonymization Training

Anonymization training is mainly comprised of a minimax optimization of utility loss and privacy loss.

Temporal-distinctiveness Objective as Utility Weakly-supervised video anomaly detection methods leverage temporal information to help localize anomalies. [28, 17, 42, 6] show the positive effect of extracted feature separability along the temporal dimension. In order to adopt the SPAct anonymization framework to anomaly detection problems, we utilize a temporal-distinctiveness objective in the utility branch. We design a temporal-distinctive triplet loss, which increases the agreement between temporally-aligned clips of the same video and increases dissimilarity between representations of the clips which are temporally misaligned. For anchor clip $\mathbf{x}_t^{(i)}$, we obtain the positive clip from the exact same timestamp, but with a differently augmented version denoted as $\tilde{\mathbf{x}}_t^{(i)}$. Whereas, the negative clip is obtained from different timestamp $\mathbf{x}_{t'}^{(i)}$, where $t' \neq t$. This triplet of clips are passed through the utility model f_T to achieve features denoted as $\mathbf{z}_t^{(i)}$, $\tilde{\mathbf{z}}_t^{(i)}$, and $\mathbf{z}_{t'}^{(i)}$. The mathematical expression for the proposed temporal-distinctive triplet loss can be expressed as follows:

$$\mathcal{L}_D^{(i)} = \max\{d(\mathbf{z}_t^{(i)}, \tilde{\mathbf{z}}_t^{(i)}) - d(\mathbf{z}_t^{(i)}, \mathbf{z}_{t'}^{(i)}) + \mu, 0\} \quad (4)$$

where $d(u_j, v_j) = \|\mathbf{u}_j - \mathbf{v}_j\|_2$ is Euclidean distance between two vectors $\mathbf{u}$ and $\mathbf{v}$, and μ is the controllable margin hyperparameter to determine how far to push and pull features in the latent space.

We utilize this loss along with the standard cross-entropy action classification loss shown in the following equation:

$$\mathcal{L}_{CE}^{(i)} = -\sum_{c=1}^{N_C} \mathbf{y}_c^{(i)} \log \mathbf{p}_c^{(i)} \quad (5)$$

where N_C is the total number of action classes of $\mathbb{D}_{action}$, $\mathbf{y}_c^{(i)}$ is the ground-truth, and $\mathbf{p}_c^{(i)}$ is prediction vector by utility model f_T.

Adding both temporal-distinctiveness (Eq. 4) and action classification objective to our utility branch, our overall utility loss can be expressed as follows

$$\mathcal{L}_T = \mathcal{L}_{CE} + \omega * \mathcal{L}_D \quad (6)$$

where ω hyperparameter is the weight of temporal-distinctive triplet loss with respect to cross-entropy loss.

Privacy (i.e. budget) Loss $\mathcal{L}_B$ We utilize the same self-supervised privacy loss from [11], which removes the private information by minimizing the agreement between the frames of the same video.

Minimax Optimization After reformulating the utility loss, we use the minimax optimization process similar to [11]. It is a two-step iterative process that minimizes the utility loss and at the same time increases budget loss $\mathcal{L}_B$. At the end of this optimization, we obtain the learned anonymization function (f_A) and utility video encoder (f_T).

3.4. Privacy-preserved Anomaly Detection Training

In order to detect whether videos from $\mathbb{D}_{anomaly}$ dataset is anomalous or not, we utilize the current state-of-the-art

technique, Magnitude-Contrastive Glance-and-Focus Network (MGFN) [6]. Similar to other recent works in anomaly detection, MGFN requires fixed features for each video from a pre-trained video encoder for anomaly detection training.

Feature Extraction In our privacy-preserving case, we can not use $\mathbb{D}_{anomaly}$ directly for the feature extraction from the video encoder. We first anonymize each video (X^i) of the dataset through the learned anonymization function f_A to get an anonymized set of the dataset. For the feature extraction, we utilize the learned utility video encoder f_T. We denote this extracted set of anonymized features as $\mathbb{F}_{anomaly} = \{\, f_T(f_A(X^i)) \mid \forall X^i \in \mathbb{D}_{anomaly} \,\}$.

Optimizing for Anomaly Detection MGFN anomaly detection is comprised of 4 main losses: (1) a standard sigmoid cross-entropy loss L_{sce} for snippet classification accuracy, (2) a temporal smoothing loss L_{ts} [40] to encourage smoothness between feature representations of consecutive segments, (3) a sparsity term L_{sp} [40] to discourage false positive anomalies, and (4) a novel magnitude contrastive loss L_{mc} to learn scene-adaptive feature distributions across videos, all which help to train a model f_{AD}.

The training loss used in MGFN is compounded in the following equation:

$$L_{AD} = L_{sce} + \lambda_1 L_{ts} + \lambda_2 L_{sp} + \lambda_3 L_{mc} \tag{7}$$

where $\lambda_1 = \lambda_2 = 1$, and $\lambda_3 = 0.001$. f_{AD} outputs frame-level anomaly scores, which are used to calculate a final ROC AUC and AP for evaluation.

3.5. Algorithm

Let's consider the models f_A, f_T, f_B, f_{AD} are parameterized by θ_A, θ_T, θ_B, and f_{AD}, respectively. All training steps of our framework can be put together in a sophisticated form of algorithm 1.

Algorithm 1: TeD-SPAD Framework

Inputs:
  Datasets: $\mathbb{D}_{action}$, $\mathbb{D}_{anomaly}$
  #Epochs: max_anon_epoch, max_ad_epoch
  Learning Rates: α_{AD}, α_B, α_T
  Hyperparameters: μ, ω
Output: θ_{AD}, θ_A

Model Initialization:
Initialize θ_T with Kinetics400 weights [1];
Initialize θ_B with SimCLR ImageNet weights [5].
Initialize $\theta_A \leftarrow \theta_A - \alpha_A \nabla_{\theta_A}(\mathcal{L}_{L1}(\theta_A))$(Ref. Eq 3)

Anonymization Training:
for $e_0 \leftarrow 1$ **to** max_anon_epoch **do**
  Step-1
  $\theta_A \leftarrow \theta_A - \alpha_A \nabla_{\theta_A}(\mathcal{L}_T(\theta_A, \theta_T) - \omega L_B(\theta_A, \theta_B))$
  Step-2
  $\theta_T \leftarrow \theta_T - \alpha_T \nabla_{\theta_T}(\mathcal{L}_T(\theta_T, \theta_A))$, (Ref. Eq 6)
  $\theta_B \leftarrow \theta_B - \alpha_B \nabla_{\theta_B}(\mathcal{L}_B(\theta_B, \theta_A))$.
end

Feature Extraction on $\mathbb{D}_{anomaly}$:
$\mathbb{F}_{anomaly} = \{\, f_T(f_A(X^i)) \mid \forall X^i \in \mathbb{D}_{anomaly} \,\}$

Privacy-Preserved Anomaly Detection Training:
for $e_0 \leftarrow 1$ **to** max_ad_epoch **do**
  $\theta_{AD} \leftarrow \theta_{AD} - \alpha_{AD} \nabla_{\theta_{AD}}(L_{AD}(\theta_{AD}, \mathbb{F}_{anomaly}))$
end

4. Experiments

4.1. Datasets

UCF-Crime [40] is the first large-scale weakly supervised video anomaly detection dataset. It contains 1,900 videos totaling 128 hours of untrimmed CCTV surveillance footage from a variety of different scenes. The videos contain 13 crime-based anomalies such as Arrest, Fighting, and Shoplifting in real world scenes. There are video-level labels indicating what anomaly is contained in each video, with the test set having frame-level anomaly annotations to evaluate performance.

XD-Violence [45] is currently the largest weakly supervised video anomaly detection dataset, totaling 217 hours of untrimmed video. All of its anomaly categories are related to violence, and each of the 4,754 videos contains audio along with video, making it useful for multi-modal anomaly detection techniques. The videos are gathered from various types of cameras, movies, and games, resulting in a unique blend of scenes for increased difficulty.

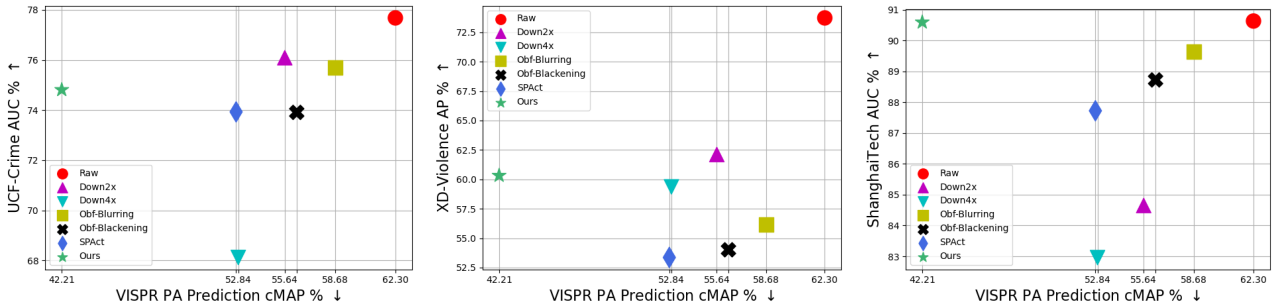
ShanghaiTech [26] is a medium-scale anomaly detection dataset containing videos covering 13 different scenes with various types of anomalies. The dataset contains pixel-level anomaly annotations which are not used in this work. While it was published as an unsupervised anomaly detection dataset, we use the weakly supervised rearrangement proposed by [51].

VISPR [29] is an image dataset labelled with 68 privacy-related attributes, including skin color, gender, hair color, clothes, etc. It provides a multi-class classification problem for us to evaluate privacy preservation on. The VISPR split we use for evaluation along with training details can found in Supp.Sec.B.

UCF101 [39] is a very common dataset in action recognition, and its relative simplicity makes it practical for learning an anoymization model on, demonstrated in [11].

4.2. Implementation Details

Network Architecture Details f_A is a UNet [35] model that transforms raw input frames into anonymized frames. I3D [1] is used for f_T, to first learn anonymized action classification, then to extract de-identified features. A ResNet-



(a) UCF-Crime (b) XD-Violence (c) ShanghaiTech

Figure 3: Trade-off plots between anomaly detection benchmarks [40, 45, 26] AUC and VISPR [29] privacy attribute prediction cMAP for different privacy preserving methods. Optimal trade-off point is top left of plot (higher AD performance, lower PA prediction ability).

50 [15] is our f_B model during the anonymization training. f_{AD} is an unmodified MGFN [6], which consists of a shortcut convolution (SCC), a self-attentional convolution (SAC), and a Feed-Forward-Network.

Training Process Details First we do anonymization training for 80 epochs. Adam [20] optimizer is used for all models with a learning rate is 1e-4 with a batchsize of 8. Loss weight $\omega = 0.1$ and margin $\mu = 1$ in a default setting. We train the MGFN [6] model using default hyperparameters. f_B evaluation uses a batch size of 32 and a base learning rate of 1e-3, which follows a linear warmup and a step-based scheduler that drops by a factor of 1/5 upon loss stagnation.

Input Details For all experiments, we crop each image to a scale of 0.8, then resize to input resolution 224×224. Clips consist of 16 frames are sampled from a random start with a skip-rate of 2. For anonymization training we utilize standard augmentations like random erase, random crop, horizontal flip and random color jitter. To maintain temporal consistency, augmentations are applied equally to every frame within each clip.

Feature Extraction Given an unmodified input video $\mathbf{X^i} \in \mathbb{D}_{anomaly}$, we first extract S clips, where S is the amount of non-overlapping 16-frame clips in $\mathbf{X}$. We sequentially pass each clip $S^{(j)}$ through first our anonymizer f_A, then our feature extractor f_T. Extracted features $f(\mathbf{X}^i)$ have the shape S x C, where C is the dimensionality of the feature vector. Specifically, features are extracted following the average pooling after the *mix_5c* layer of I3D and have a dimensionality of 2048.

Evaluation Protocol and Performance Metrics To evaluate the learned anonymization function f_A, we follow standard protocols of cross-dataset evaluation [11, 46]. In this protocol, testing videos of $\mathbb{D}_{anomaly}$ are anonymized by f_A, and frame-level predictions are obtained through the f_T and f_{AD}. The calculated ROC AUC is used to evaluate performance on UCF-Crime and ShanghaiTech, and AP is used for XD-Violence. A higher AUC is considered a more accurate anomaly localization. In order to evaluate the privacy leakage, the learned f_A is utilized to anonymize the privacy dataset $\mathbb{D}_{privacy}$ to train and evaluate a target privacy model f'_B. Privacy leakage is measured in terms of the performance of the target f'_B on the test set of $\mathbb{D}_{privacy}$. Since the privacy dataset is multi-label, the privacy leakage is measured in terms of mean average precision averaged across classes (cMAP).

More implementation details in Supp. Sec C.

4.3. Privacy Preserving Baselines

We run well-known self-supervised privacy preservation techniques for video anomaly detection. In order to maintain a fair comparison across methods, we utilize the exact same network architectures and training process.

Downsampling Baselines For `Downsample-2x` and `Downsample-4x`, we simply resize the input frames to a lower resolution by a factor of 2 (112×112) and 4 (56×56).

Object-Detector Based Obfuscation Baselines Obfuscation techniques are based on first detecting the person, followed by removing (i.e blackening) or blurring them. Both obfuscation techniques use MS-COCO [24] pre-trained Yolo [33] object detector to obtain bounding boxes for `person` object class. We utilize YOLOv5[1] implementation with yolov5x as backbone. The detected bounding boxes are assigned to pixel value 0 for the `Blackening`-based baseline. For the `Blurring`-based baselines, a Gaussian filter with kernel $k = 13$ and variance $\sigma = 10$ is utilized.

SPAct [11] Baseline We utilize official implementation[2]. For a fair comparison with our method, we utilize the exact same utility model I3D and privacy model ResNet-50.

4.4. Evaluation on Benchmark Anomaly Datasets

We compare prior privacy-preserving methods to our method on 3 well-known anomaly detection benchmark

[1] https://github.com/ultralytics/yolov5
[2] https://github.com/DAVEISHAN/SPAct

Feature Extraction Model	VISPR Privacy cMAP (%)	VISPR Privacy Reduction (%)
Kinetics400 pretrained	63.15	0.0
SPAct [11] Anonymized	55.60	11.96
Our Anonymized	52.30	17.18

Table 1: Quantitative evidence of action classification model features leaking privacy. Features from each model used to predict privacy attributes. Red indicates higher privacy leakage.

datasets. Since privacy-preservation deals with both utility (i.e. anomaly) and privacy, we show results in form of a trade-off plot as presented in Fig. 3. Compared to the prior best method [11], our method achieves is able to remove 19.9% more privacy with a slightly better utility score (1.19%). This strongly supports our claim that promoting temporal distinctiveness during anonymization better aligns with anomaly detection tasks. Numeric data behind Fig. 3 plots can be found in Supp. Sec. D.

4.5. Qualitative Results

Figure 4 shows visual examples of the model outputs in different videos. We note that to the human eye, it is difficult to tell what is going on in each video. Looking closely reveals the main subjects and some outlines of the surroundings, which intuitively may be enough for anomaly detection. None of the private attributes of the human subjects are visible, and therefore cannot be used to make unfair decisions. See Supp.Sec.D for more qualitative results.

4.6. Evidence for Privacy Leakage at Feature-level

In anomaly feature representation learning, the anomaly detection algorithms do not directly work with the input videos, the videos are first passed through an action classifier to compute features. Even though the action recognition model sees the original videos, it is not certain whether the private information gets passed to the features. In order to confirm this, we create a simple fully connected network to predict VISPR private attributes, in the same fashion that we evaluate privacy for our other experiments. We stack the same VISPR image 16 times to create a video clip, then extract the clip features through f_T. Our baseline uses unmodified input images passed through a pretrained Kinetics400 [18] I3D [1] model, with the other experiments using a paired anonymizer and tuned I3D model. Detailed explanations of this process are found in Supp.Sec.C. Through experimentation, we find that *the action classifier latent features do in fact leak private information, therefore this private information gets passed into the anomaly detector*. Table 1 demonstrates empirical evidence of this.

4.7. Ablation Study

Effect of different utility losses $\mathcal{L}_T$ We study the effect of different utility losses during the anonymization process on the final privacy vs utility anomaly detection performance in Table 2. From `row-(a,b)`, we see a clear reduction in privacy at no cost of anomaly performance; which demonstrates the effectiveness of our proposed temporal-distinctive triplet loss $\mathcal{L}_D$.

	Utility Loss during Anonymization ($\mathcal{L}_T$)	VISPR Privacy cMAP(%)(↓)	UCF-Crimes Anomaly AUC(%)(↑)
(a)	$\mathcal{L}_{CE}$	52.71	73.93
(b)	$\mathcal{L}_{CE} + \mathcal{L}_D$	**42.21**	**74.81**
(c)	$\mathcal{L}_{CE} + \mathcal{L}_I$	45.64	69.52

Table 2: Ablation with different utility losses during the anonymization process. Bold indicates best trade-off.

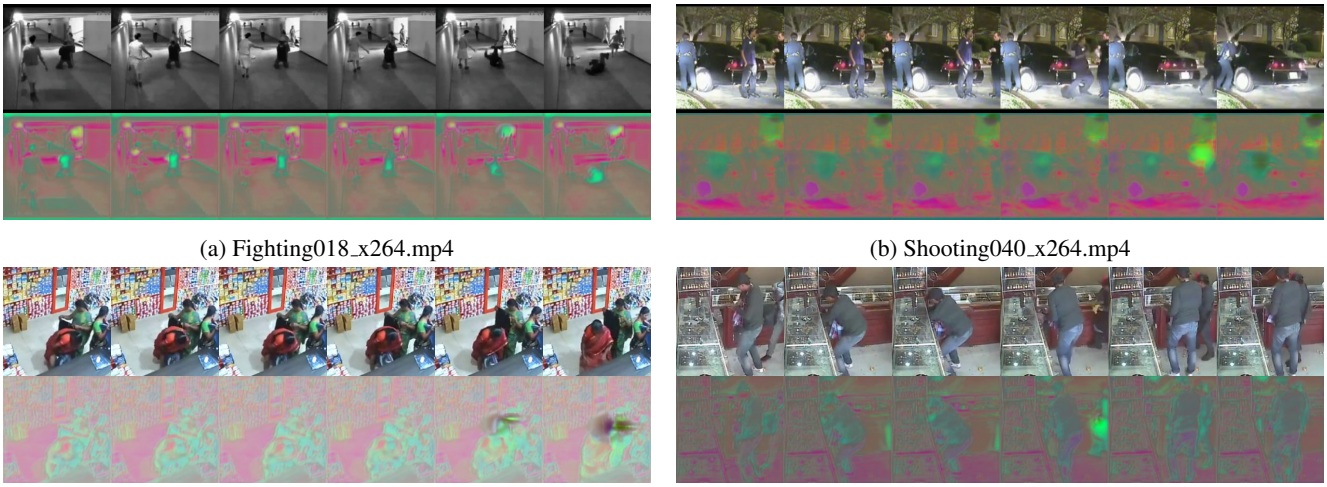
Triplet Temporal Loss Weight ω	VISPR Privacy cMAP(%)(↓)	UCF-Crime Anomaly AUC(%)(↑)
0	52.71	73.93
0.01	53.84	72.53
0.1	**42.21**	**74.81**
1.0	55.26	70.56
10.0	51.67	69.27

Table 3: Comparison of using different loss weights of the temporal-distinctive triplet loss during the anonymization process. The margin hyperparameter of the temporal triplet loss for each experiment was 1. Bold indicates best trade-off.

To this extent, we also implement the contrary objective to our $\mathcal{L}_D$ which promotes temporal-invariance $\mathcal{L}_I$ in the utility branch of the anonymization training. We implement this using well-known self-supervised works [32, 10]. From `row-(c)` we see that temporal-invariance objective is not well-suited for anomaly detection utility tasks and results in a significant drop of **6%**. We provide extensive experiments with $\mathcal{L}_I$ and its explanation in Supp. Sec. D.

Relative Weightage of $\mathcal{L}_D$ Here we test the effect of changing the weight of the additional temporal-distinctive triplet loss. Table 3 shows that weighting the loss at 0.1 achieves our best results of 32.25% relative increase in privacy with only a 3.69% reduction in utility performance. Without the enforced temporal distinctiveness, the utility model is limited by the quality of the reconstructed anonymized videos. The improper weighting of the temporal loss interferes with the classifier ability of the model, which can also harm the anonymization process. This suggests that action recognition loss $\mathcal{L}_{CE}$ is still the important utility task for anomaly detection performance.

Effect of the margin in $\mathcal{L}_D$ The proposed $\mathcal{L}_D$ temporally-distinctive triplet loss uses a margin hyperparameter μ to allow for adjustments to the contrastive distance in the la-



(a) Fighting018_x264.mp4 (b) Shooting040_x264.mp4

(c) Shoplifting048_x264.mp4 (d) Robbery023_x264.mp4

Figure 4: Here we show four examples of clips from the UCF-Crime [40] dataset. For each pairing, the top row is the unmodified video, and the bottom is the clip after being passed through the anonymizer.

tent feature space. The only requirement is that $\mu > 0$. The intuition here is that a larger margin enforces greater feature spacing. As Table 4 shows, we empirically found that setting $\mu = 1$ gives us the most robust results. A lower margin results in less temporally-distinct representations, which makes distinguishing between normal and anomalous features more difficult. On the other hand, increasing the $\mu = 2$ results in more difficult temporal triplet loss (i.e a very high temporal distinctiveness) which may not align well with anomaly detection task.

Triplet Temporal Loss Margin μ	**VISPR** Privacy cMAP(%)(↓)	**UCF-Crime** Anomaly AUC(%)(↑)
0.5	49.78	62.12
1.0	**42.21**	**74.81**
2.0	58.86	67.30

Table 4: Comparison of using different hyperparameter margins of the temporal-distinctive triplet loss during the anonymization process. Bold indicates default setup.

Effect of temporal sampling in $\mathcal{L}_D$ Proposed triplet loss forms negative from the clip $\mathbf{x}_{t'}^{(i)}$ of a different timestamp t'. Distance between the timestamp of negative and anchor clip $t - t'$ is an important aspect to define temporal distinctiveness. We perform experiments with various distances as shown in Table 5. In our default setting, we use random distance as shown in the first row. From the second row, we can say that a smaller distance leads to a better anomaly score with a slight degradation in protecting privacy. At the same time, the third row suggests that enforcing temporal-distinctiveness at higher distances leads to better privacy protection but at the cost of anomaly performance. This distance hyperparameter may be used as a tuning parameter to get the different operating points of privacy vs anomaly trade-off.

Negative Clip Distance	**VISPR** Privacy cMAP(%)(↓)	**UCF-Crime** Anomaly AUC(%)(↑)
Random	**42.21**	**74.81**
8	46.34	76.12
32	28.69	70.97

Table 5: Comparison of enforcing set clip sampling distance during the anonymization process. The margin hyperparameter of the temporal triplet loss for each experiment was 1.

5. Conclusion

In this paper, we highlight the importance of privacy, a previously neglected aspect of video anomaly detection. We present TeD-SPAD, a framework for applying Temporal Distinctiveness to Self-supervised Privacy-preserving video Anomaly Detection. Ted-SPAD demonstrates the effectiveness of using a temporal-distinctive triplet loss while anonymizing an action recognition model, as it enhances feature representation temporal distinctiveness, which complements the downstream anomaly detection model. By effectively destroying spatial private information, we remove the model's ability to use this information in its decision-making process. As a future research direction, this framework can be extended to other tasks, such as spatio-temporal anomaly detection. The anonymizing encoder-decoder may also be made more powerful with techniques using recent masked image modeling. It is our hope that this work contributes to the development of more responsible and unbiased automated anomaly detection systems.

References

[1] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, Feb. 2018. arXiv:1705.07750 [cs]. 4, 5, 7

[2] Chun-Fu Chen, Shaohan Hu, Zhonghao Shi, Prateek Gulati, Bill Moriarty, Marco Pistoia, Vincenzo Piuri, and Pierangela Samarati. Mass: Multi-attribute selective suppression. *arXiv preprint arXiv:2210.09904*, 2022. 2

[3] Haoyang Chen, Xue Mei, Zhiyuan Ma, Xinhong Wu, and Yachuan Wei. Spatial–temporal graph attention network for video anomaly detection. *Image and Vision Computing*, 131:104629, Mar. 2023. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, June 2020. arXiv:2002.05709 [cs, stat]. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 4, 5

[6] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection, Nov. 2022. arXiv:2211.15098 [cs]. 2, 3, 4, 5, 6

[7] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456, June 2011. ISSN: 1063-6919. 2

[8] 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

[9] 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, Sept. 2015. 2

[10] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. TCLR: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, June 2022. 7

[11] Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. Spact: Self-supervised privacy preservation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3, 4, 5, 6, 7

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[13] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2021. 3

[14] Giacomo Giorgi, Wisam Abbasi, and Andrea Saracino. Privacy-Preserving Analysis for Remote Video Anomaly Detection in Real Life Environments. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 13(1):112–136, Mar. 2022. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, Dec. 2015. arXiv:1512.03385 [cs]. 6

[16] Carlos Hinojosa, Miguel Marquez, Henry Arguello, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. PrivHAR: Recognizing Human Actions From Privacy-preserving Lens, June 2022. arXiv:2206.03891 [cs]. 2

[17] Chao Huang, Chengliang Liu, Jie Wen, Lian Wu, Yong Xu, Qiuping Jiang, and Yaowei Wang. Weakly Supervised Video Anomaly Detection via Self-Guided Temporal Discriminative Transformer. *IEEE Transactions on Cybernetics*, pages 1–14, 2022. Conference Name: IEEE Transactions on Cybernetics. 3, 4

[18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset, May 2017. arXiv:1705.06950 [cs]. 7

[19] Junho Kim, Young Min Kim, Yicheng Wu, Ramzi Zahreddine, Weston A. Welge, Gurunandan Krishnan, Sizhuo Ma, and Jian Wang. Privacy-Preserving Visual Localization with Event Cameras, Dec. 2022. arXiv:2212.03177 [cs]. 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] B. Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos. *Journal of Imaging*, 4(2):36, Feb. 2018. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. 2

[22] Ming Li, Jun Liu, Hehe Fan, Jia-Wei Liu, Jiahe Li, Mike Zheng Shou, and Jussi Keppo. STPrivacy: Spatio-Temporal Tubelet Sparsification and Anonymization for Privacy-preserving Action Recognition, Jan. 2023. arXiv:2301.03046 [cs]. 2

[23] Shuo Li, Fang Liu, and Licheng Jiao. Self-Training Multi-Sequence Learning with Transformer for Weakly Supervised Video Anomaly Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):1395–1403, June 2022. 3

[24] Tsung-Yi Lin, Michael Maire, Serge J. 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*, 2014. 6

[25] 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

[26] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 5, 6

[27] Zelun Luo, Jun-Ting Hsieh, Niranjan Balachandar, Serena Yeung, Guido Pusiol, Jay Luxenberg, Grace Li, Li-Jia Li, N Lance Downing, Arnold Milstein, and Li Fei-Fei. Com-

puter Vision-Based Descriptive Analytics of Seniors' Daily Activities for Long-Term Health Monitoring. 2

[28] Hui Lv, Chuanwei Zhou, Chunyan Xu, Zhen Cui, and Jian Yang. Localizing Anomalies from Weakly-Labeled Videos. *IEEE Transactions on Image Processing*, 30:4505–4515, 2021. arXiv:2008.08944 [cs]. 3, 4

[29] 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. 2, 5, 6

[30] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning Memory-guided Normality for Anomaly Detection, Mar. 2020. arXiv:2003.13228 [cs]. 2

[31] Oluwatoyin P. Popoola and Kejun Wang. Video-Based Abnormal Human Behavior Recognition—A Review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):865–878, Nov. 2012. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2

[32] 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. 7

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, June 2016. ISSN: 1063-6919. 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

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. arXiv:1505.04597 [cs]. 5

[36] 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

[37] Hitesh Sapkota and Qi Yu. Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection, Mar. 2022. arXiv:2203.12840 [cs]. 3

[38] Rakshith Shetty, Mario Fritz, and Bernt Schiele. Adversarial Scene Editing: Automatic Object Removal from Weak Supervision, June 2018. arXiv:1806.01911 [cs, stat]. 2

[39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, Dec. 2012. arXiv:1212.0402 [cs]. 5

[40] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, Salt Lake City, UT, June 2018. IEEE. 1, 2, 5, 6, 8

[41] Qiyue Sun and Yang Yang. Unsupervised video anomaly detection based on multi-timescale trajectory prediction. *Computer Vision and Image Understanding*, 227:103615, Jan. 2023. 2

[42] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning, Aug. 2021. arXiv:2101.10030 [cs] version: 3. 3, 4

[43] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *ECCV*, 2022. 3

[44] Peng Wu and Jing Liu. Learning Causal Temporal Relation and Feature Discrimination for Anomaly Detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021. Conference Name: IEEE Transactions on Image Processing. 3

[45] Peng Wu, jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 5, 6

[46] 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 (TPAMI)*, 2020. 1, 2, 3, 6

[47] Adam Yala, Victor Quach, Homa Esfahanizadeh, Rafael G. L. D'Oliveira, Ken R. Duffy, Muriel Médard, Tommi S. Jaakkola, and Regina Barzilay. Syfer: Neural Obfuscation for Private Data Release, Jan. 2022. arXiv:2201.12406 [cs]. 2

[48] Hongchun Yuan, Zhenyu Cai, Hui Zhou, Yue Wang, and Xiangzhi Chen. TransAnomaly: Video Anomaly Detection Using Video Vision Transformer. *IEEE Access*, 9:123977–123986, 2021. Conference Name: IEEE Access. 2

[49] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting Completeness and Uncertainty of Pseudo Labels for Weakly Supervised Video Anomaly Detection, Dec. 2022. arXiv:2212.04090 [cs]. 3

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

[51] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li. Graph Convolutional Label Noise Cleaner: Train a Plug-and-play Action Classifier for Anomaly Detection, Mar. 2019. arXiv:1903.07256 [cs]. 3, 5

[52] Yi Zhu and Shawn Newsam. Motion-Aware Feature for Improved Video Anomaly Detection, July 2019. arXiv:1907.10211 [cs, eess]. 3