Invisibility Cloak: Hiding Anomalies in Videos via Adversarial Machine Learning Attacks

Hamza Karim
University of South Florida
hamzakarim@usf.edu

Yasin Yilmaz*
University of South Florida
yasiny@usf.edu

Abstract

Video anomaly detection (VAD) plays a crucial role in various fields, providing an invaluable tool for enhancing security, safety, and operational efficiency. Video anomaly detection systems are designed to identify irregular patterns, unusual behaviors, or unexpected events within a given video stream. Among these, weakly supervised VAD (wVAD) systems have gained significant popularity due to their ability to leverage anomalous video samples (i.e., weak labels), which can be easily obtained in many applications unlike anomalous frame samples (i.e., fully labelled data). The superior performance of wVAD systems compared to unsupervised VAD methods makes wVAD systems particularly attractive in real-world applications of security surveillance and content moderation for online video streaming platforms. The potential use of wVAD systems in such critical applications also raises concerns regarding potential adversarial machine learning attacks. Adversaries may exploit vulnerabilities within these systems to evade detection, posing significant risks to the security and integrity of the system. This study explores the vulnerabilities of wVAD systems by comprehensively analyzing the weaknesses of these systems under a white-box setting. We propose a metric for quantifying the efficacy of such attacks and show that practical attacks can achieve up to 99% success rate in hiding anomalies.

1. Introduction

Video anomaly detection (VAD) has attracted significant research interest since the advent of deep learning models. Acquisition of inclusive labelled data for VAD systems has always been a challenge, leading to increased attention towards unsupervised VAD and weakly supervised VAD methodologies. Weakly supervised VAD (wVAD) systems have demonstrated higher detection performance compared to unsupervised approaches for trained anomaly types, primarily attributed to their capacity to incorporate video-level weak labels during training and utilize large pre-trained video feature extractors. Nevertheless, it is known that such pre-trained video models are susceptible to adversarial attacks. Thus, this study aims to assess the efficacy of adversarial attacks targeting recent state-of-the-art wVAD systems to evaluate their robustness.

Adversarial machine learning attacks have recently garnered significant attention in various domains, especially in computer vision tasks such as image and action recognition. Following the groundwork by Goodfellow et al., which demonstrated the vulnerability of image recognition models to adversarial perturbations in the FGSM [5] paper, subsequent research has extended the examination of adversarial attacks to diverse vision-related challenges, including video action recognition. However, exploring adversarial attacks remains an open problem within the domain of video anomaly detection. This paper focuses on developing adversarial attacks on popular wVAD systems to investigate their robustness. Similar to action recognition, attacks can be white-box or black-box and targeted or untargeted. In a white-box setting, the model and its parameters are assumed known while in a black-box attack, the model under attack is assumed unknown. Despite the model assumption, white-box attacks still have practical value since an attack designed on a surrogate model can be transferable to many other models, especially those based on a similar CNN or transformer architecture. In a targeted scenario, the adversarial attack's objective may involve hiding the anomalies by reducing the true positive rate or increasing the false positive rate to compromise the integrity of the system. Conversely, in an untargeted scenario, the aim is to render the system generally unreliable while preserving the content of the video to the greatest extent possible.

In this research, we focus on developing targeted white-box attacks against state-of-the-art wVAD systems. We propose an attack model that can be trained to attack a wVAD system and attack the system *in real-time* during inference by generating restricted perturbations in the video content.

*This work was supported by U.S. National Science Foundation (NSF) under grant 2040572.

To demonstrate the effectiveness of our attack architecture, we test the model on two popular wVAD datasets, UCF-Crime [18] and XD-Violence [25]. Our contributions can be summarized as follows:

- A novel deep learning architecture for attacking wVAD systems, which achieves up to 100% success in targeted attacks.
- Investigating the robustness of state-of-the-art wVAD systems against adversarial attacks.
- Investigating the attack transferability across different models and datasets.
- New metrics for measuring the efficacy of targeted adversarial attacks on VAD systems.

2. Related Works

2.1. Video Anomaly Detection

Recent research has shown an increasing interest in VAD systems. VAD is studied in two settings, unsupervised and weakly supervised. While the unsupervised VAD algorithms train only on nominal data, in weakly supervised VAD (wVAD), the training set includes a number of anomalous videos without any detailed annotation about the nature of the involved anomaly (e.g., when, what kind, how many). This enables wVAD methods to focus on the anomaly types important for the use case, as a result of which wVAD methods get better at detecting those anomaly types compared to unsupervised VAD methods.

The widespread availability of anomalous samples with video-level labels has motivated research in wVAD systems. Sultani et al. [18] introduced a deep multiple-instance learning framework for anomalous segment detection. RTFM [20] focused on a feature magnitude and a multi-scale temporal scenario to select top-k segments for abnormality assessment. The same paper also proposed the now widely adopted, multi-scale temporal network (MTN) for feature aggregation. S3R [24] utilizes dictionary-based self-supervised learning with MTN network to generate enhanced pseudo-anomalous video features. MGFN [2] introduces a Glance-and-Focus module and Magnitude Contrastive loss to enhance normal and abnormal feature separability. More recently, REWARD [7] focused on real-time detection of anomalies by end-to-end training a transformer-based feature extractor.

2.2. Attacks to Video Understanding Systems

Adversarial attacks to video action recognition have been studied extensively under a white-box setting. Nathan et al. [6] proposed a white-box attack on two stream flow-based video action recognition models. Wei et al. [23] introduced an optimization algorithm based on $L2, 1$ norm to calculate sparse adversarial perturbations. Their focus was on networks employing a CNN+RNN architecture to analyze perturbation propagation properties. Li et al. [11] developed an offline universal perturbation using a GAN-based model. This perturbation was applied to unseen input in real-time video recognition models. Chen et al. [3] used the concept of appending adversarial frames in a video stream to attack popular video models such as C3D [21] and I3D [1]. Over-the-air flicker attack [15] investigated the robustness of various video models by producing flickering perturbations across video frames.

In recent years, researchers have begun to investigate vulnerabilities in surveillance systems through physical attacks. Much research has been focused on attacking human body and facial detectors. Thys et al. [19] introduced a method to create adversarial patches aimed at fooling automated surveillance cameras by iteratively adjusting them based on objectness and class loss metrics from the YOLO detector. Notably, researchers in [26] successfully integrated these patches into t-shirt designs to deceive modern human detectors. However, Xu et al. [28] found that the effectiveness of these adversarial t-shirts diminishes in real-world situations due to the dynamic nature of human movement and body flexibility. To address these challenges, they suggested using Thin Place Spline (TPS) mapping to better model the possible deformations encountered by moving and non-rigid objects.

Similarly, [17] demonstrated that specifically engineered spectacle frames possess the capability to deceive state-of-the-art facial recognition systems. Additionally, Yin et al. [29] advocate for a makeup blending technique intended to enhance the efficacy of adversarial perturbations through a fine-grained learning attack strategy. Furthermore, Komkov et al. [8] devised a rectangular paper sticker intended for head-wear as a means to subvert the ArcFace [4] facial recognition model.

Beyond physical attacks, the reliability of VAD systems is underexplored. Mumcu et al. [13] examined how cyber-attacks, like WiFi de-authentication, degrade VAD performance through video stream disruptions (e.g., lower resolution, freezing, and speed changes). Xie et al. [27] used a U3D video action recognition black-box attack on a MIL-based system, showing significant performance drops. However, no comprehensive study has yet addressed adversarial attacks designed to stealthily target modern wVAD systems using representative datasets.

3. Threat Model

An adversary may find an incentive to target a wVAD system such as a surveillance infrastructure or a content moderation system. The motivation to attack a surveillance system can be to mask potentially harmful activities (e.g., terrorist attack, robbery, assault, etc.), whereas

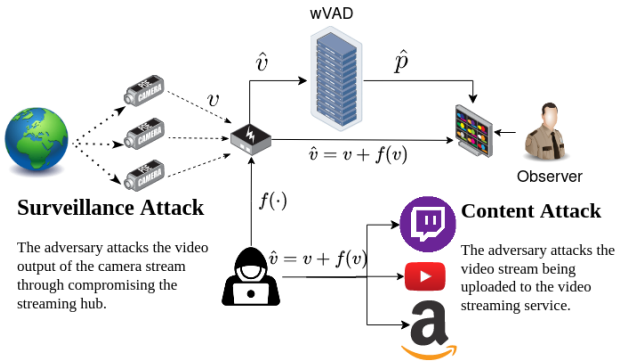

Figure 1. Threat Model: In case of surveillance attack, the adversary may deploy malware into the (IoT) streaming hub, thereby compromising the integrity of the system. Conversely, in the scenario of a content attack, the adversary might directly introduce noise into the content.

in the content hiding attack, the attacker tries to evade detection algorithms utilized by popular streaming platforms like YouTube, Facebook, and Instagram. Figure 1 illustrates the attack process for both scenarios. The attacker designs a function $f(\cdot)$, perturbs the original video v with the attack vector $f(v)$ so that the perturbed video $\hat{v} = v + f(v)$ is received by the wVAD system. The attacker's goal is to mislead the wVAD system to a wrong decision $p(\hat{v})$.

We now discuss the assumptions under which the attack operates. It is assumed that the attacker can manipulate the video sent from the camera to the streaming hub (e.g., Blink sync module, Ring alarm base station). Considering the typically low level of security in such IoT hubs, in particular the community edge drivers running on these hubs [16], we assume that the attacker can insert a code into the hub to slightly modify the video in a stealthy way that would not alert a human observer yet alter the decision of wVAD system. Such a carefully designed attack would look like normal noise or glitch in the streaming system. In the case of evading content moderation detectors, the attacker has direct access to the original video, which can either be offline or online depending on the specific use cases. The attacker's purpose is to bypass the detector while maintaining the illegal content. Since we consider a white box setting, we also assume the attacker knows the parameters of the target model. Additionally, we also assume the attacker has access to the data domain used to train the target model.

It is reasonable to assume that one or more of these assumptions may not hold in practical scenarios, hence we also evaluate the efficacy of our attack under conditions where discrepancies may exist between the target and surrogate model, as well as between the training and test data domains. In Section 5.3, we show that the attack is transferable if both the surrogate and target model are CNN-based. However, the attack is not transferable if one of them is transformer-based and the other is CNN-based.

4. Practical Attacks to wVAD Systems

Our goal is to train a algorithm $f(\cdot)$ to hide anomalies (false negative attack). This attack is supposed to generate noise masks that are concentrated to only specific pixels in each frame with reasonably small perturbation magnitude. Code and a demo are available in Supplementary.

4.1. Attack Design

Adherent to the wVAD literature, we first divide each video into T segments, each containing F frames. Given each segmented video $v \in \mathbb{R}^{T\times C\times F\times H\times W}$, where C is the number of channels, H and W are the height and width of each frame, an anomaly detector $Q(v)$ processes it to compute the anomaly probability $p(v)$. The attack function $f(\cdot)$ operates on the input video $v \in \mathbb{R}^{T\times C\times F\times H\times W}$

Unlike iterative white-box attacks like FGSM [5], IGSM [9], and PGD [12] , which require gradient computations for each input, we train a generative model that can attack anomaly detectors in real-time without gradient calculations, while offering some transferability across data domains. Wang et al. [22] and Naseer et al. [14] proposed a framework for such attacks on image recognition systems. These approaches utilize a generative model composed of an encoder-decoder network with residual connections, producing an output the same size as the input. In this study, this method based on 2D convolutions is referred to as the CLOAK-2D attack. Drawing inspiration from this, we extend the approach to videos by employing 3D convolution layers instead of 2D convolution layers to capture temporal information. We propose a simple video encoder-decoder attack network, comprising of three down-sampling 3D convolution layers and three up-sampling 3D transposed convolution layers, with residual connections between the encoder and decoder layers. We term this method as CLOAK-3D attack.

To enhance the CLOAK-3D attack to better suit the VAD setting, we introduce an attention-based temporal aggregation block to the CLOAK-3D architecture, enabling the network to learn temporal dependencies across the T segments. This enhanced method is referred to as the CLOAK-3D-A attack, which is discussed in detail in Section 4.3. In the subsequent sections, we will test and compare the efficacy of these three attack methods when targeting three state-of-the-art wVAD systems.

Figure 2 illustrates the architecture of the image-based CLOAK-2D and the video-based CLOAK-3D attack networks. The produced perturbation mask $f(v)$ is added to the clean video v to obtain the corresponding adversarial video $\hat{v} = v + f(v)$. For attack training, our proposed approach leverages the loss gradients derived from passing a video v through the attack function $f(\cdot)$ and passing the

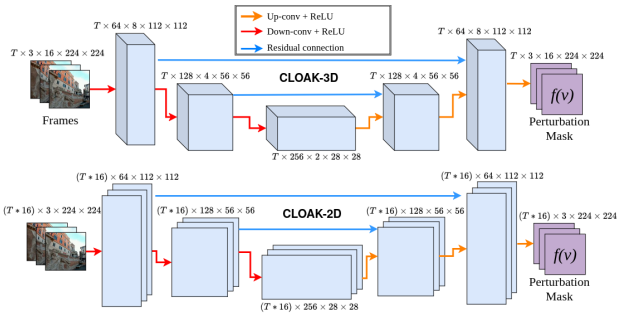

Figure 2. CLOAK-3D attack uses 3D convolution and transpose convolution layers to upsample and downsample video segments, leveraging the temporal and spatial dimensions simultaneously. In contrast, a CLOAK-2D attack utilizes 2D convolution and transpose convolution layers on individual frames within each video segment, focusing solely on the spatial dimensions of the frames.

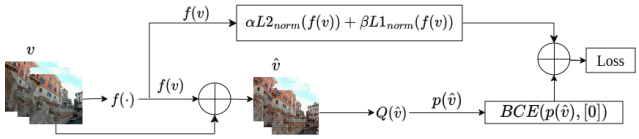

Figure 3. Training pipeline

resultant perturbed video $\hat{v}$ through the anomaly detector $Q(\hat{v})$. The gradients are back-propagated through the attack model $f(v)$ as illustrated in Figure 3. Note that after each back pass, the anomaly detector $Q(\hat{v})$ gradients are reset.

In the wVAD setting, we have access to two sets of data subsets, one consisting of nominal videos and the other consisting of anomalous videos. Training of CLOAK is exclusively conducted on anomalous videos, significantly reducing both training time and computational resources. The target label for each video is set to zero, hence the Binary Cross-Entropy (BCE) loss is computed as follows:

$$BCE = -\frac{1}{N}\sum_{n=1}^{N}\log(1 - p(\hat{v}_n))$$

where N represents the number of anomalous videos in the dataset and $p(\hat{v}_n)$ denotes the predicted anomaly score of perturbed video n. Back-propagating the gradient of this loss through the attack model encourages the generation of perturbations that drive $p(\hat{v})$ towards zero.

4.2. Regularization

We quantify the distortion introduced by the perturbation matrix $f(v)$ in the spatio-temporal domain. This metric will be constrained in order for the perturbation to be imperceptible to the human observer while remaining adversarial. In the literature, the widely adopted way of restricting perturbations is to clamp the maximum perturbation strength between $-\epsilon$ and ϵ, where ϵ is a small arbitrary number indicating the tolerance parameter. In our experiments, we impose a similar restriction using a fixed value for ϵ. In addition to that, we also regularize the perturbations using both the L2-Norm and the L1-Norm of the perturbation matrix $f(v)$. The objective of the L1-Norm is to concentrate perturbations on significant frames and pixels across frames for each video segment, while the L2-Norm is used to limit the magnitude of these perturbations.

In the final loss,

$$\text{Loss} = BCE + \alpha\|f(v)\|_2 + \beta\|f(v)\|_1, \tag{1}$$

α and β are constants used to control L2 and L1 regularization. Equation (1) encapsulates the combined loss used to train the proposed CLOAK attacks, incorporating BCE along with L2 and L1 regularization terms.

4.3. Temporal Aggregation

Temporal feature aggregation has been a fundamental technique for enhancing the performance of wVAD systems [20], [2], [24]. The popular temporal aggregation method, known as multi-scale temporal network (MTN) [20], aggregates features across T segments, providing a broader context to the VAD system beyond a single segment. We employ this concept to design an attack network capable of learning temporal dependencies across T segments using an attention mechanism. Figure 4 depicts the architecture underlying the CLOAK-3D-A method.

Building on the work of Tian et al. (2021) [20], we aim to aggregate the feature maps produced by the encoder across the temporal dimension T. MTN uses 1D convolutions on the feature vector extracted by passing each segment through a feature extractor such as I3D. We extend this concept to the 3D feature maps generated by the encoder, applying four 3D convolutions with varying dilation values to achieve a larger receptive field. The outputs are denoted as $X1$, $X2$, $X3$, and $X4$.

$X4$ is fed into the QKV-attention block and converted into query, key, and value feature maps. These feature maps are then flattened to compute attention between the vectors. Notably, the transpose of the query and value is taken instead of the key. This transposed attention computes the attention for each individual feature in the feature map across the temporal dimension T. The attention block's output is reshaped and concatenated with $X1$, $X2$, and $X3$. These four concatenated feature maps are subsequently fused through a final convolution layer to produce an output $\hat{X}$ with dimensions similar to the input, which are then added together before being passed to the decoder. We later analyze the benefits of this aggregation in Section 5.2

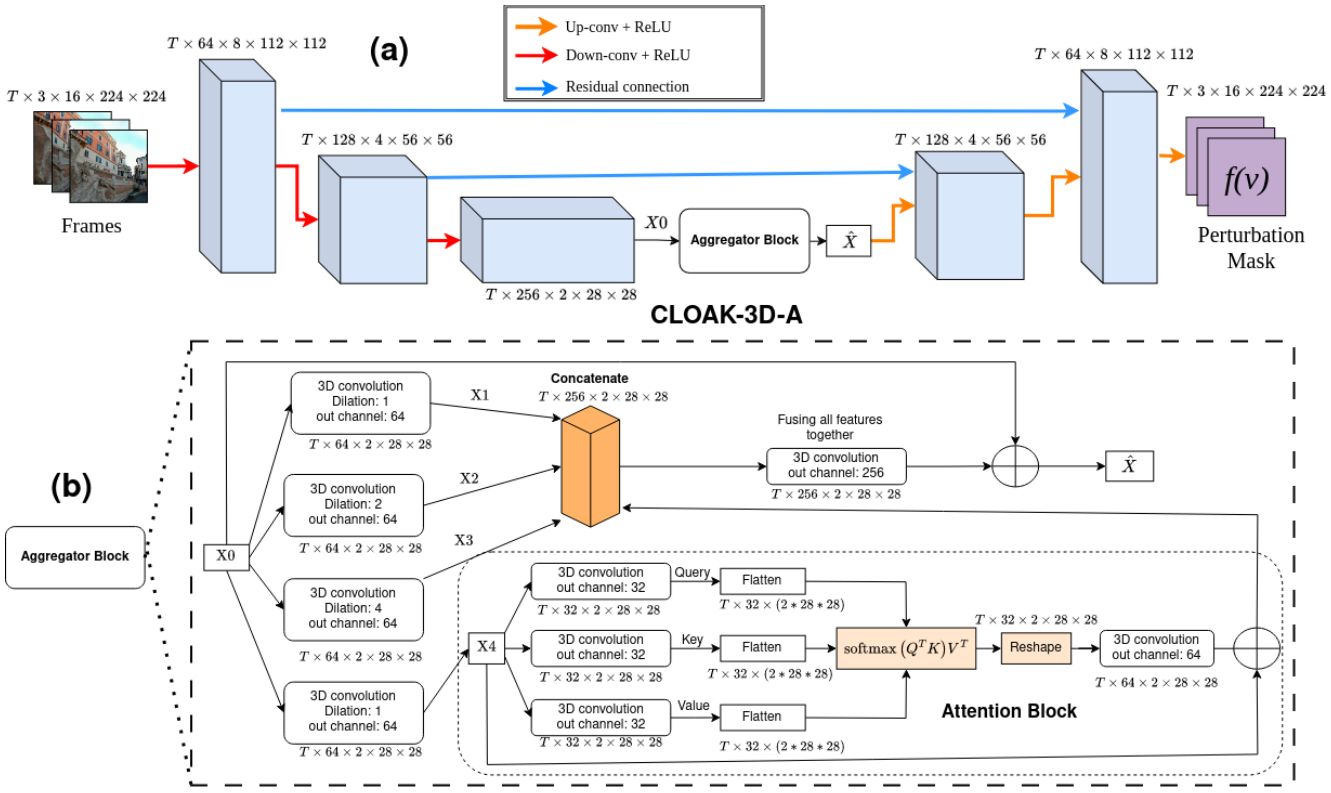

Figure 4. (a) The CLOAK-3D-A attack adds an attention-based aggregation block between the output of the encoder and the input of the decoder. (b) Detailed representation of the Aggregator block.

5. Experiments

We test the efficacy of our proposed attack model on three state-of-the-art wVAD algorithms on two datasets UCF-Crime [18] and XD-Violence [25]. Most contemporary research in wVAD uses I3D as the pre-trained feature extractor, however, recent methods have also used more modern transformer-based architectures. We test our model on two state-of-the-art methods that use I3D as their backbone, namely S3R [24] and MGFN [2]. We also test our approach on a more recent method REWARD [7] that uses Uniformer [10] as its backbone. In Section 5.3, we also discuss the transferability of our proposed attack to have a deeper understanding of the attack's efficacy. For all our experiments we set $T = 32$, $H = W = 224$, $\alpha = 10^{-4}$ and $\beta = 10^{-7}$. We choose the value of ϵ based on the normalized standard deviation for each frame. The normalization used by all standard video and image feature extractors use 0.45 and 0.225 as mean and standard deviation. We set $\epsilon = 0.1125$, which is half of the standard deviation of the normalized frames. Different values of ϵ are also tested in Section 6.1. We use an Adam optimizer with the learning rate and weight decay set to 10^{-4}.

5.1. Metrics

The common practice in the literature is to use the Area Under the ROC Curve (ROC_{AUC}) metric for evaluating performance on the UCF-Crime dataset and the Average Precision (AP) metric, which is the area under the precision-recall curve, on the XD-Violence dataset. However, when measuring the success of a targeted attack on an anomaly detection system, the drop in these metrics alone is not sufficient. The focus for evaluating attack success should be on the percentage of missed true positives due to the attack. To address this, we adopt a methodology wherein the success rate of targeted attacks is evaluated across a spectrum of thresholds, spanning from 1% to 99% false positive rate (FPR) tolerance for each system. At each threshold, the objective is to diminish the total count of true positives, hence hiding anomalies. Consequently, we define the success rates for the attack as follows:

$$\text{Success Rate} = 1 - \frac{\text{No. of true positives after attack}}{\text{No. of true positives before attack}}. \quad (2)$$

We propose a novel metric aimed at quantifying the efficacy of attacks on VAD systems. After plotting the success rate (as defined in Equation 2) vs. FPR, the resultant Area Under the Curve (AUC) for each of these plots is defined

as $Attack_{AUC}$. This new metric provides a more suitable measure of attack efficacy since the drop in ROC_{AUC} normalizes the errors due to the attack by the number of actual positives/negatives while the focus should be on the number of detected positives/negatives by the model before the attack.

5.2. Results

In this section, we discuss the results of our proposed attack. In Figures 5 and 6, we present the $Attack_{AUC}$ curves for all three methods explained in Section 4 on both UCF-Crime and XD-Violence datasets. We represent our results with the following annotation:

$$M_1 \rightarrow M_2 | D_1 \rightarrow D_2,$$

where M_1 is the surrogate model $Q(v)$ used to train our attack model $f(\cdot)$, and M_2 is the target model under attack. D_1 is the data domain used to train $f(\cdot)$ while D_2 is the test domain. We also provide perturbation budget examples on a sample test video sourced from the UCF-Crime dataset.

The first model under attack is S3R. The method uses I3D as its pre-trained feature extractor and implements a methodology based on dictionary-based self-supervised learning to produce en-normal and de-normal features. Additionally, the model integrates the MTN network to produce temporally enhanced features, thereby enabling model to learn pseudo-anomalous video features.

Figure 5(a) and 6(a) illustrate that at lower FPR values, the success rate of a false negative (i.e., anomaly hiding) attack approaches 100%. However, as the FPR increases due to a threshold decrease towards zero, the success rate for false negative attack diminishes since the target model generously raises anomaly alarms. Note that, in practical settings, target models typically operate under stringent FPR constraints, where the proposed attacks achieve perfect success rates. We see that the CLOAK-3D-A attack significantly outperforms the CLOAK-2D and CLOAK-3D attacks on both UCF-Crime and XD-Violence datasets.

Additionally, Figure 7(a) shows a sample of the perturbed frame produced by the CLOAK-3D-A attack on a video sample from UCF-Crime. In our experiments, we observe that these perturbations are limited to specific regions within a frame and are barely visible to the human eye. Figure 7(b) demonstrates the effectiveness of the attack on the sample video as we observe that the anomaly scores predicted by the target model after the attack are close to zero throughout the video.

The second model under consideration is MGFN. The method introduces a Glance-and-Focus module alongside Magnitude Contrastive loss, utilizing feature magnitudes to improve the differentiation between normal and abnormal features. This method also uses I3D as its pre-trained feature extractor and MTN for temporal aggregation of features. Figure 5(b) and 6(b) shows a similar trend to attacks on S3R. While in this case CLOAK-2D performs at par with CLOAK-3D, both fall behind the attention-based aggregation variant. Figure 8 shows, similar to S3R, the stealthy perturbations are able to reduce the anomaly scores to near zero.

Table 1. Transferabilty analysis between CNN-based detectors.

	$Attack_{AUC}$%		
	CLOAK-3D-A	CLOAK-3D	CLOAK-2D
S3R→MGFN|UCF→UCF	**92.49**	89.64	82.99
S3R→MGFN|XD→XD	24.32	22.87	**25.98**
MGFN→S3R|UCF→UCF	**85.75**	85.44	80.3
MGFN→S3R|XD→XD	**51.18**	44.23	33.5

The third model targeted in our study is REWARD. This method focuses on real-time anomaly detection and proposes to train a transformer-based feature extractor end-to-end using pseudo-labels generated through $k$nn-distances. Given its end-to-end nature, this model amalgamates the anomaly detector and the feature extractor into a unified model.

The attack plots for REWARD follow similar trends to the previous attacks, as illustrated in Figure 5(c) and 6(c). Since REWARD's transformer architecture fundamentally different, the convolution-based attack is less effective compared to the convolution-based S3R and MGFN, as expected. We also observe in case of XD-Violence, the performance of the CLOAK-3D sustains much of its performance at higher FPR values when compared to the other variants. However, at lower FPR values, CLOAK-3D-A is still the best performing attack, hence resulting in a higher $Attack_{AUC}$. It is also seen in Figure 9 that the perturbations produced by CLOAK-3D-A to attack REWARD display a more grid like pattern due to the patch-based processing of transformer. While REWARD has more variation in the predicted anomaly scores throughout the video, the CLOAK-3D-A attack brings all anomaly scores down to approximately zero, as shown in Figure 9(b).

5.3. Transferability

A major practical concern for white-box attacks is transferability of the attacks across various parameters, such as model architecture and data domain. In this section, we investigate the effectiveness of our attack under two conditions: model mismatch such that $M_1 \neq M_2$ and data domain mismatch such that $D_1 \neq D_2$.

5.3.1 Model Mismatch Analysis:

We first analyze the efficacy of attacks when the surrogate and target models differ ($M_1 \neq M_2$).

Table 1 presents the transferability results of the attack between the S3R and MGFN, exhibiting varying degrees of effectiveness. We see that in most cases the aggregated

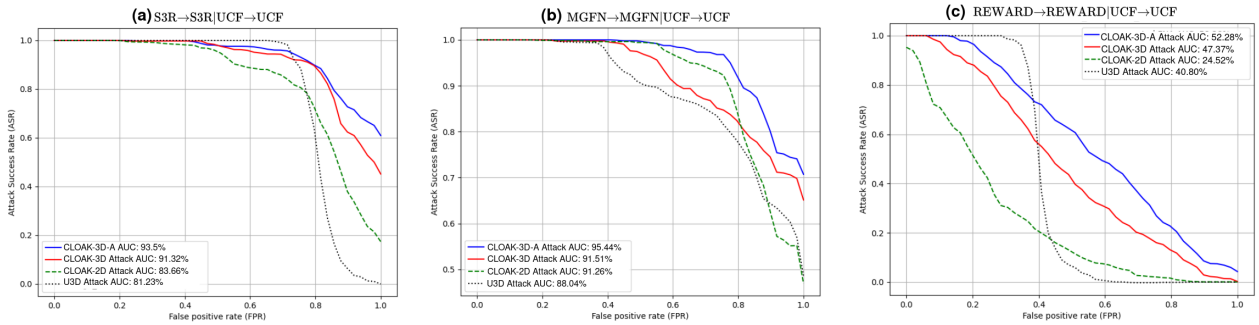

Figure 5. Attack success rate plots for attacks against (a)S3R, (b)MGFN, and (c)REWARD when trained and tested on UCF-Crime.

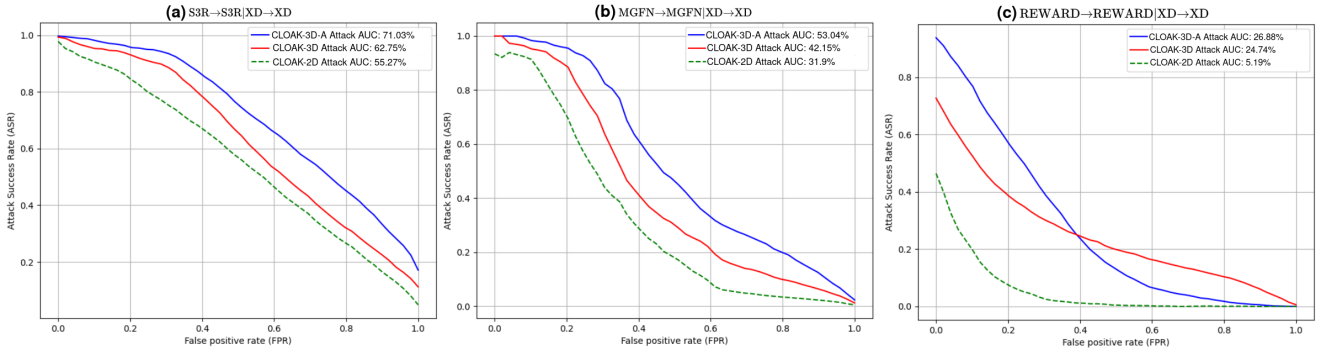

Figure 6. Attack success rate plots for attacks against (a)S3R, (b)MGFN, and (c)REWARD when trained and tested on XD-Violence.

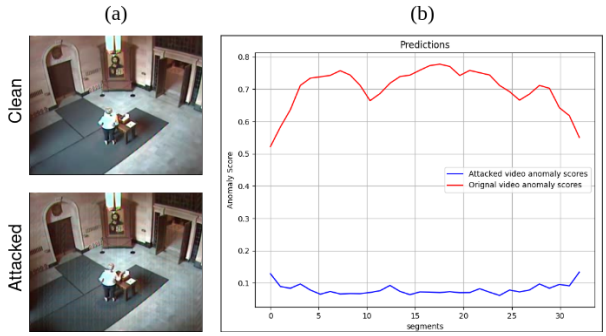

Figure 7. (a) Sample video attacked stealthily using the CLOAK-3D-A model under the condition S3R→S3R|UCF→UCF. (b) Anomaly score predicted by the target model is significantly reduced after the CLOAK-3D-A attack, indicating successful anomaly hiding.

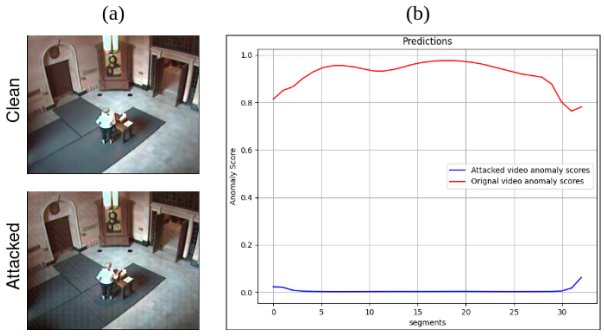

Figure 8. (a) Sample video attacked stealthily using the CLOAK-3D-A model under the condition MGFN→MGFN|UCF→UCF. (b) Anomaly score predicted by the target model is significantly reduced after the CLOAK-3D-A attack, indicating successful anomaly hiding.

variant outperforms the other attack methods except in the case of surrogate model being S3R and the target model being MGFN, trained and tested on XD-Violence. The overall performance of all models drop with CLOAK-2D being marginally better than CLOAK-3D-A.

The use of a transformer-based Uniformer-32 feature extractor in REWARD distinguishes it from anomaly detectors employing a CNN-based I3D feature extractor. Consequently, it is rational to anticipate diminished transferabil-

Table 2. Transferability analysis between transformer-based and CNN-based anomaly detectors.

	$Attack_{AUC}$%		
	CLOAK-3D-A	CLOAK-3D	CLOAK-2D
S3R→REWARD|UCF→UCF	3.02	1.28	11.13
MGFN→REWARD|UCF→UCF	1.03	1.22	7.5
REWARD→S3R|UCF→UCF	5.43	0.78	2.75
REWARD→MGFN|UCF→UCF	9.97	2.73	5.77

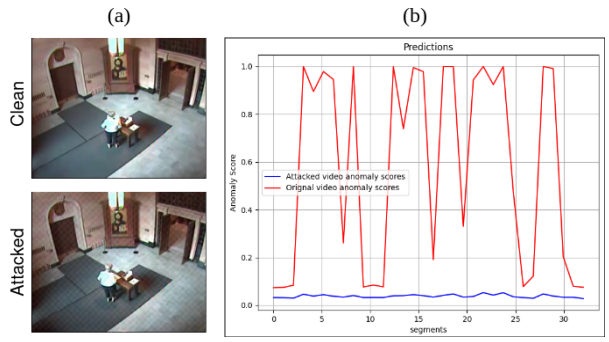

Figure 9. (a) Sample video attacked stealthily using the CLOAK-3D-A model under the condition REWARD→REWARD|UCF→UCF. (b) Anomaly score predicted by the target model is significantly reduced after the CLOAK-3D-A attack, indicating successful anomaly hiding.

Table 3. Transferability analysis between UCF-Crime and XD-Violence data domains.

	$Attack_{AUC}\%$		
	CLOAK-3D-A	CLOAK-3D	CLOAK-2D
S3R→S3R|UCF→XD	62.81	**69.65**	60.19
S3R→S3R|XD→UCF	**74.22**	60.94	47.35
MGFN→MGFN|UCF→XD	**39.01**	29.07	26.96
MGFN→MGFN|XD→UCF	**66.92**	55.33	40.46
REWARD→REWARD|UCF→XD	**24.33**	13.46	6.31
REWARD→REWARD|XD→UCF	**46.4**	25.41	16.69

ity of attacks between these distinct frameworks. Table 2 underscores this argument as we see limited transferability across such disparate architectures. We see that CLOAK-2D sustains some of its performance, however observing the plots for these results available in supplementary material, we see that the drop in the success rate is steep even at lower FPR values.

5.3.2 Data Domain Mismatch Analysis:

Often times the adversary may not have the exact data domain the anomaly detector was trained on. Hence, it is useful to investigate the attack efficacy across data domain. We investigate for each of the three models such that $D_1 \neq D_2$.

Table 3 provides evidence of some transferability across the data domain. Especially when the attack model is trained on a more diverse large-scale dataset XD-Violence and tested on UCF-Crime. It is observed that in most cases CLOAK-3D-A maintains over 30% performance with the exception of REWARD→REWARD|UCF→XD, where all attacks see a significant drop in performance.

6. Ablation Study

In this section, we present an ablation study aimed at controlling and limiting the perturbations produced by the function $f(v)$ by controlling the value of ϵ. We train and test all three attack models with three different values of ϵ on the UCF-Crime dataset. We also test the effect of α and β on the attack performance.

Table 4. Attack performance vs. attack budget values ϵ for all attacks tested on UCF-Crime with no model or data mismatch.

		$Attack_{AUC}\%$		
Target	Attack Method	$\epsilon = 0.05625$	$\epsilon = 0.1125$	$\epsilon = 0.225$
S3R	CLOAK-3D-A	**90.08**	**93.5**	96.87
	CLOAK-3D	86.45	91.32	96.03
	CLOAK-2D	70.5	83.66	**96.93**
MGFN	CLOAK-3D-A	**80.52**	**95.44**	95.45
	CLOAK-3D	73.22	91.52	96.03
	CLOAK-2D	65.05	91.26	**96.08**
REWARD	CLOAK-3D-A	**31.32**	**58.28**	**77.9**
	CLOAK-3D	24.77	47.37	67.57
	CLOAK-2D	23.15	24.52	38.8

Table 5. Effect of regularization parameters α and β on CLOAK-3D-A attack under the setting S3R→S3R|UCF→UCF.

$Attack_{AUC}\%$		
$\beta = 1e{-}7$		
$\alpha = 1e{-}5$	$\alpha = 1e{-}4$	$\alpha = 1e{-}3$
92.52	93.5	86.77
$\alpha = 1e{-}4$		
$\beta = 1e{-}8$	$\beta = 1e{-}7$	$\beta = 1e{-}6$
96.87	93.5	74.12

6.1. Effect of ϵ on attack performance

Table 4 demonstrates that, in most cases, the CLOAK-3D-A significantly outperforms the other methods. However, it is noteworthy that as the budget is increased to a higher value, the performance of the three attack methods tend to converge. Nonetheless, in the case of the REWARD detector, we observe that even at a higher value of ϵ, there remains a considerable performance gap between each method.

6.2. Effect of α and β

Table 5 shows that increasing both α and β (i.e., more regularized perturbations) has negetive impact on performance, as expected. Specifically, β, the coefficient of L1 loss, seems to have a larger effect on performance. Sparser perturbations affect more than reduced magnitude. It is evident that $\alpha = 10^{-5}$ produces similar performance compared to $\alpha = 10^{-4}$ due to clipping perturbations by ϵ. Hence, further reduction in α may not have a significant impact on performance.

7. Conclusion

We conducted a comprehensive study on the vulnerability of modern wVAD systems to adversarial attacks. In a white-box setting, attackers can train a neural network with encoder-decoder architecture to create real-time stealthy perturbations. Testing three wVAD methods on two benchmark datasets, we found that 3D convolutions and attention allow attackers to hide anomalies (e.g., robbery, violence) in surveillance videos, especially at lower false alarm rates. Our transferability analysis showed the attack remains effective under model mismatch if both models are convolution-based and under data mismatch if the training set is larger and more diverse than the test set.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[2] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 387–395, 2023. 2, 4, 5

[3] Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Qi Tian. Appending adversarial frames for universal video attack. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3199–3208, 2021. 2

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 3

[6] Nathan Inkawhich, Matthew Inkawhich, Yiran Chen, and Hai Li. Adversarial attacks for optical flow-based action recognition classifiers. *arXiv preprint arXiv:1811.11875*, 2018. 2

[7] Hamza Karim, Keval Doshi, and Yasin Yilmaz. Real-time weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6848–6856, 2024. 2, 5

[8] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 819–826. IEEE, 2021. 2

[9] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 3

[10] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022. 5

[11] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy Chowdhury, and Ananthram Swami. Adversarial perturbations against real-time video classification systems. *arXiv preprint arXiv:1807.00458*, 2018. 2

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9), 2017. 3

[13] Furkan Mumcu, Keval Doshi, and Yasin Yilmaz. Adversarial machine learning attacks against video anomaly detection systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 206–213, 2022. 2

[14] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[15] Roi Pony, Itay Naeh, and Shie Mannor. Over-the-air adversarial flickering attacks against video recognition networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 515–524, 2021. 2

[16] Samsung. Third-party edge drivers are not maintained or reviewed by smartthings. https://developer.smartthings.com/docs/devices/hub-connected/enroll-in-a-shared-channel. Accessed: 2024-11-07. 3

[17] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 22(3):1–30, 2019. 2

[18] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 2, 5

[19] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2

[20] 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. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986, 2021. 2, 4

[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[22] Zhibo Wang, Hongshan Yang, Yunhe Feng, Peng Sun, Hengchang Guo, Zhifei Zhang, and Kui Ren. Towards transferable targeted adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20534–20543, 2023. 3

[23] Xingxing Wei, Jun Zhu, Sha Yuan, and Hang Su. Sparse adversarial perturbations for videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8973–8980, 2019. 2

[24] 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 *European Conference on Computer Vision*, pages 729–745. Springer, 2022. 2, 4, 5

[25] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer, 2020. 2, 5

[26] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 1–17. Springer, 2020. 2

[27] Shangyu Xie, Han Wang, Yu Kong, and Yuan Hong. Universal 3-dimensional perturbations for black-box attacks on video recognition systems. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1390–1407. IEEE, 2022. 2

[28] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 665–681. Springer, 2020. 2

[29] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. *arXiv preprint arXiv:2105.03162*, 2021. 2