

Over-the-Air Adversarial Flickering Attacks against Video Recognition Networks

Roi Pony^{1*}, Itay Naeh^{2*}, Shie Mannor^{1,3}

¹Department of Electrical Engineering, Technion Institute of Technology, Haifa, Israel

² Rafael - Advanced Defense Systems Ltd., Israel

³ Nvidia Research

roipony@gmail.com, itay@naeh.us, shie@technion.ac.il

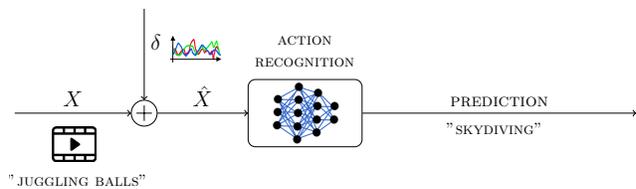
Abstract

Deep neural networks for video classification, just like image classification networks, may be subjected to adversarial manipulation. The main difference between image classifiers and video classifiers is that the latter usually use temporal information contained within the video. In this work we present a manipulation scheme for fooling video classifiers by introducing a flickering temporal perturbation that in some cases may be unnoticeable by human observers and is implementable in the real world. After demonstrating the manipulation of action classification of single videos, we generalize the procedure to make universal adversarial perturbation, achieving high fooling ratio. In addition, we generalize the universal perturbation and produce a temporal-invariant perturbation, which can be applied to the video without synchronizing the perturbation to the input. The attack was implemented on several target models and the transferability of the attack was demonstrated. These properties allow us to bridge the gap between simulated environment and real-world application, as will be demonstrated in this paper for the first time for an over-the-air flickering attack.

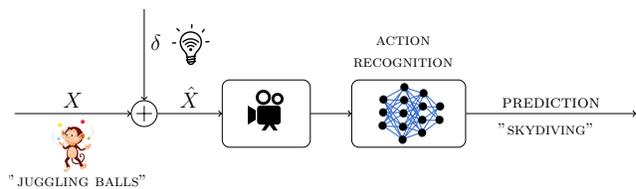
1. Introduction

In recent years, Deep Neural Networks (DNNs) have shown phenomenal performance in a wide range of tasks, such as image classification [12], object detection [18], semantic segmentation [22] etc. Despite their success, DNNs have been found vulnerable to adversarial attacks. Many works [27, 5, 17] have shown that a small (sometimes imperceptible) perturbation added to an image, can make a given DNNs prediction false. These findings have raised

*Equal contribution



(a) Diagram of a Flickering Adversarial Attack in a simulated environment (digital).



(b) Diagram of an Over-the-Air Flickering Adversarial Attack in the real-world (physical).

Figure 1: Top figure shows the attack diagram in the digital domain performed by adding a uniform RGB perturbation to the attacked video. Bottom figure shows the modeling of the digitally-developed attack into the real-world by transmitting the perturbation in the scene using a smart RGB led bulb.

many concerns, particularly for critical systems such as face recognition systems [25], surveillance cameras [24], autonomous vehicles, and medical applications [16]. In recent years most of the attention was given to the study of adversarial patterns in images and less in video action recognition. Only in the past two years works on adversarial video attacks were published [33, 8, 34, 10], even though DNNs have been applied to video-based tasks for several years, in particular video action recognition [2, 32, 4]. In video action recognition networks temporal information is of the essence in categorizing actions, in addition to per-frame im-

age classification. In some of the proposed attacks the emphasis was, beyond adversarial categorization, the sparsity of the perturbation. In our work, we consider adversarial attacks against video action recognition under a white-box setting, with an emphasis on the imperceptible nature of the perturbation in the spatio-temporal domain to the human observer and implementability of the generalized adversarial perturbation in the real-world. We introduce flickering perturbations by applying a uniform RGB perturbation to each frame, thus constructing a *temporal* adversarial pattern. Unlike previous works, in our case sparsity of the pattern is undesirable, because it helps the adversarial perturbation to be detectable by human observers for its unnatural pattern, and to image based adversarial perturbation detectors for the exact same reason. The adversarial perturbation presented in this work does not contain any spatial information on a single frame other than a constant offset. This type of perturbation often occurs in natural videos by changing lighting conditions, scene changes, etc. In this paper, we aim to attack the video action recognition task [11]. For the targeted model we focus on the I3D [2] model (Specifically we attack the RGB stream of the model, rather than on the easier to influence optical flow stream) based on InceptionV1 [26] and we expand our experiments to additional models from [29]. The attacked models trained on the Kinetics-400 Human Action Video Dataset [11].

In order to make the adversarial perturbation unnoticeable by human observers, we reduce the thickness and temporal roughness of the adversarial perturbation, which will be defined later in this paper. In order to do so we apply two regularization terms during the optimization process, each corresponds to a different effect of the perceptibility of the adversarial pattern. In addition, we introduce a modified adversarial-loss function that allows better integration of these regularization terms with the adversarial loss.

We will first focus on the I3D [2] network and introduce a flickering attack on a single video and present the trade-off between the different regularization terms. We construct universal perturbations that generalize over classes and achieve 93% fooling ratio. Another significant feature of our proposed method is time invariant perturbations that can be applied to the classifier without synchronization. This makes the perturbation relevant for real world scenarios, since frame synchronization is rarely possible. We show the effectiveness of the flickering attack on other models [29] and the inter-model transferability, and finally demonstrate the over-the-air flickering attack in a real world scenario for the first time. A diagram of the digital attack and the over-the-air attack pipelines are shown in Figure 1. The main contributions of this work are:

- A methodology for developing flickering adversarial attacks against video action recognition networks that incorporates a new type of regularization for affecting

the visibility of the adversarial pattern.

- A universal time-invariant adversarial perturbation that does not require frame synchronization.
- Adversarial attacks that are transferable between different networks.
- Adversarial attacks that are implementable using temporal perturbations.

The paper is organized as follows: We briefly review related work and present the flickering adversarial attack. Then we show experimental results and the generalization of the attack. Finally, we present real world examples of the flickering adversarial attacks, followed by conclusions and future work. We encourage the readers to view the attack videos¹, over-the-air scene-based attack videos², and over-the-air universal attack videos³. Our code can be found in the following repository⁴.

2. Related Work

2.1. Video Action Recognition

With deep Convolutional Neural Networks (CNNs) achieving state-of-the-art performance on image recognition tasks, many works propose to adapt this achievement to video-based computer vision tasks. The most straightforward approach for achieving this is to add temporally-recurrent layers such as LSTM [21] models to traditional 2D-CNNs. This way, long-term temporal dependencies can be assigned to spatial features [31, 23]. Another approach implemented in C3D [9, 28, 30] extends the 2D CNNs (image-based) to 3D CNNs (video-based) kernels and learns hierarchical spatio-temporal representations directly from raw videos. Despite the simplicity of this approach, it is very difficult to train such networks due to their huge parameter space. To address this, [2] proposes the Inflated 3D CNN (I3D) with inflated 2D pre-trained filters [20]. In addition to the RGB pipeline, optical flow is also useful for temporal information encoding, and indeed several architectures greatly improved their performance by incorporating an optical-flow stream [2]. [29] demonstrated the advantages of 3D CNNs over 2D CNNs within the framework of residual learning, proposing factorization of the 3D convolutional filters into separate spatial and temporal components.

2.2. Adversarial Attack on Video Models

The research of the vulnerability of video-based classifiers to adversarial attacks emerged only in the past years.

¹https://bit.ly/Flickering_Attack_videos

²https://bit.ly/Over_the_Air_scene_based_videos

³https://bit.ly/Over_the_Air_videos

⁴https://bit.ly/Flickering_Attack_Code

The following attacks were performed under the white-box attack settings: [33] were the first to investigate a white-box attack on video action recognition. They proposed an $L_{2,1}$ norm based optimization algorithm to compute sparse adversarial perturbations, focusing on networks with a CNN+RNN architecture in order to investigate the propagation properties of perturbations. [15] generated an offline universal perturbation using a GAN-based model that they applied to the learned model on unseen input for real-time video recognition models. [19] proposed a nonlinear adversarial perturbation by using another neural network model (besides the attacked model), which was optimized to transform the input into adversarial pattern under the L_1 norm. [8] proposed both white and black box untargeted attacks on two-stream model (optical-flow and RGB), based on the original and the iterative version of FGSM [5, 13], and used FlowNet2 [7] to estimate optical flow in order to provide gradients estimation. Several black-box attacks were proposed [10, 34]. Our attack follows the white-box setting therefore those attacks are beyond the scope of this paper.

3. Flickering Adversarial Attack

The flickering adversarial attack consists of a uniform offset added to the entire frame that changes each frame. This novel approach is desirable for several reasons. First, it contains no spatial pattern within individual frames but an RGB offset. Second, this type of perturbation can easily be mistaken in some cases as changing lighting conditions of the scene or typical sensor behaviour. Third, it is implementable in the real-world using a simple LED light source.

3.1. Preliminaries

Video action recognition is a function $F_\theta(X) = y$ that accepts an input $X = [x_1, x_2, \dots, x_T] \in \mathbb{R}^{T \times H \times W \times C}$ from T consecutive frames with H rows, W columns and C color channels, and produces an output $y \in \mathbb{R}^K$ which can be treated as probability distribution over the output domain, where K is the number of classes. The model F implicitly depends on some parameters θ that are fixed during the attack. The classifier assigns the label $A_\theta(X) = \operatorname{argmax}_i y_i$ to the input X . We denote adversarial video by $\hat{X} = X + \delta$ where the video perturbation $\delta = [\delta_1, \delta_2, \dots, \delta_T] \in \mathbb{R}^{T \times H \times W \times C}$, and each individual adversarial frame by $\hat{x}_i = x_i + \delta_i$. \hat{X} is adversarial when $A_\theta(\hat{X}) \neq A_\theta(X)$ (untargeted attack) or $A_\theta(\hat{X}) = k \neq A_\theta(X)$ for a specific predetermined incorrect class $k \in [K]$ (targeted attack), while keeping the distance between \hat{X} and X as small as possible under the selected metric (e.g., L_2 norm).

3.2. Methodology

In our attack δ_i is designed to be spatial-constant on the three color channels of the frame, meaning that for each pixel in image x_i , an offset is added with the same value

(RGB). Thus, the i^{th} perturbation δ_i , which corresponds to the i^{th} frame x_i of the video, can be represented by three scalars, hence $\delta = [\delta_1, \delta_2, \dots, \delta_T] \in \mathbb{R}^{T \times 1 \times 1 \times 3}$, having in total $3T$ parameter to optimize. To generate an adversarial perturbation we usually use the following objective function

$$\operatorname{argmin}_\delta \lambda \sum_j \beta_j D_j(\delta) + \frac{1}{N} \sum_{n=1}^N \ell(F_\theta(X_n + \delta), t_n) \quad (1)$$

$$s.t \hat{x}_i \in [V_{min}, V_{max}]^{H \times W \times C}, \quad (2)$$

where N is the total number of training videos, X_n is the n^{th} video, $F_\theta(X_n + \delta)$ is the classifier output (probability distribution or logits), and t_n is the original label (in the case of untargeted attack). The first term in Equation (1) is regularization term, while the second is adversarial classification loss, as will be discussed later in this paper. The parameter λ weighs the relative importance of being adversarial and also the regularization terms. The set of functions $D_j(\cdot)$ controls the regularization terms that allows us to achieve better imperceptibility for the human observer. The parameter β_j weighs the relative importance of each regularization term. The constraint in Equation (2) guarantees that after applying the adversarial perturbation, the perturbed video will be clipped between the valid values: V_{min}, V_{max} , that represents the minimum and maximum allowed pixel intensity.

3.3. Adversarial loss function

We use a loss mechanism similar to the loss presented by C&W [1], with a minor modification. For untargeted attack:

$$\ell(y, t) = \max \left(0, \min \left(\frac{1}{m} \ell_m(y, t)^2, \ell_m(y, t) \right) \right) \quad (3)$$

$$\ell_m(y, t) = y_t - \max_{i \neq t} (y_i) + m, \quad (4)$$

where $m > 0$ is the desired margin of the original class probability below the adversarial class probability. A more detailed explanation of the motivation in defining the above loss function is found in the supplementary material.

3.4. Regularization terms

We quantify the distortion introduced by the perturbation δ with $D(\delta)$ in the spatio-temporal domain. This metric will be constrained in order for the perturbation δ to be imperceptible to the human observer while remaining adversarial. Unlike previously published works on adversarial patches in images, in the video domain imperceptible may reference thin patches in gray-level space or slow changing patches in temporal frame space. In contrast to previous related works [33, 34], in our case temporal sparsity is not of the essence but the unnoticability to the human observer. In

order to achieve the most imperceptible perturbation we introduce two regularization terms, each controlling different aspects of human perception.

In order to simplify the definition of our regularization terms and metrics, we define the following notations for $X = [x_1, x_2, \dots, x_T] \in \mathbb{R}^{T \times H \times W \times C}$ (video or perturbation).

Tensor p -norm:

$$\|X\|_p = \left(\sum_{i_1=1}^T \cdots \sum_{i_4=1}^C |x_{i_1 \dots i_4}|^p \right)^{1/p}, \quad (5)$$

where i_1, i_2, \dots, i_4 refer to dimensions.

Roll operator: $Roll(X, \tau)$ produce the time shifted tensor, whose elements are τ -cyclic shifted along the first axis (time):

$$Roll(X, \tau) = [x_{(\tau \bmod T)+1}, \dots, x_{(T-1+\tau \bmod T)+1}]. \quad (6)$$

1st and 2nd order temporal derivatives: We approximate the 1st and 2nd order temporal derivatives by finite differences as follows.

$$\frac{\partial X}{\partial t} = Roll(X, 1) - Roll(X, 0), \quad (7)$$

$$\frac{\partial^2 X}{\partial t^2} = Roll(X, -1) - 2Roll(X, 0) + Roll(X, 1). \quad (8)$$

3.4.1 Thickness regularization

This loss term forces the adversarial perturbation to be as small as possible in gray-level over the three color channels (per-frame), having no temporal constraint and can be related to the ‘‘thickness’’ of the adversarial pattern.

$$D_1(\delta) = \frac{1}{3T} \|\delta\|_2^2,$$

where $\|\cdot\|_2$ defined in Equation (5) with $p = 2$.

3.4.2 Roughness regularization

We introduce temporal loss functions that incorporate two different terms,

$$D_2(\delta) = \frac{1}{3T} \left\| \frac{\partial \delta}{\partial t} \right\|_2^2 + \frac{1}{3T} \left\| \frac{\partial^2 \delta}{\partial t^2} \right\|_2^2, \quad (9)$$

where $\frac{\partial \delta}{\partial t}$ and $\frac{\partial^2 \delta}{\partial t^2}$ are defined in Equations (7,8), respectively.

The norm of the first order temporal difference shown in the Equation (9) (first term) controls the difference between each two consecutive frame perturbations. This term penalizes temporal changes of the adversarial pattern. Within the

Attack	Attacked Model	Fooling ratio[%]	Thickness[%]	Roughness[%]
Single Video	I3D	100	1.0±0.5	0.83± 0.4
Single Video	R(2+1)D	93.0	2.4±1.9	2.1± 2.0
Single Class	I3D	90.2± 11.72	13.0± 3.6	10.6± 2.2
Universal	I3D	93.0	15.5	15.7
Universal	R(2+1)D	79.0	18.1	21.0
Universal	MC3	77.1	18.3	24.5
Universal	R3D	90.3	17.8	25.5
Universal Time Invariance	I3D	83.1	18.0	14.0

Table 1: Results over several types of attacks on different attacked models. Thickness and Roughness defined in Equations (10,11)

context of human visual perception, this term is perceived as ‘‘flickering’’, thus we wish to minimize it.

The norm of the second order temporal difference shown in Equation (9) (second term) controls the trend of the adversarial perturbation. Visually, this term penalizes fast trend changes, such as spikes, and may be considered as scintillation reducing term.

The weights of D_1 and D_2 will be noted by β_1 and β_2 , respectively, throughout the rest of the paper and also in the YouTube videos.

3.5. Metrics

Let us define several metrics in order to quantify the performance of our adversarial attacks.

Fooling ratio: is defined as the percentage of adversarial videos that are successfully misclassified (higher is better).

Mean Absolute Perturbation per-pixel:

$$thickness_{gl}(\delta) = \frac{1}{3T} \|\delta\|_1, \quad (10)$$

where $\|\cdot\|_1$ defined in Equation (5) with $p = 1$.

Mean Absolute Temporal-diff Perturbation per-pixel:

$$roughness_{gl}(\delta) = \frac{1}{3T} \left\| \frac{\partial \delta}{\partial t} \right\|_1. \quad (11)$$

The thickness and roughness values in this paper will be presented as percents from the full applicable values of the image span, e.g.,

$$thickness(\delta) = \frac{thickness_{gl}(\delta)}{V_{max} - V_{min}} * 100.$$

4. Experiments on I3D

4.1. Targeted Model

Our attack follows the white-box setting, which assumes the complete knowledge of the targeted model, its parameter values and architecture. The I3D [2] model for video recognition is used as target model, focused on the RGB pipeline. The adversarial attacks described in this work can

be a targeted or untargeted, and the theory and implementation can be easily adapted accordingly. The I3D model was selected for targeting because common video classification networks are based upon its architecture. Therefore, the insights derived from this work will be relevant for these networks. In the I3D configuration $T = 90, H = 224, W = 224, C = 3$, and $V_{min} = -1, V_{max} = 1$ (trained on the kinetics Dataset). Implementation details can be found in the supplementary material.

4.2. Dataset

We use Kinetics-400 [11] for our experiments. Kinetics is a standard benchmark for action recognition in videos. It contains about 275K video of 400 different human action categories (220K videos in the training split, 18K in the validation split, and 35K in the test split). For the single video attack we have developed the attacks using the validation set. In the class generalization section we trained on the training set and evaluated on the validation set. In the universal attack section we trained on the validation set and evaluated on the test set. We pre-processed the dataset by excluding movies in which the network misclassified to begin with and over-fitted entries. Each video contains 90-frame snippets.

4.3. Single Video Attack

In order to perform the flickering adversarial attacks on single videos, a separate optimization has to be solved for each video, i.e., solving Equation (1) for a single video ($N = 1$) s.t. each video have its own tailor-made δ . In our experiment we have developed different δ 's for hundreds of randomly picked samples from the kinetics validation set. The *Single Video* entry in Table 1 shows the statistics of average and standard deviation of the fooling ratio, thickness and roughness of untargeted single-video attacks, reaching 100% fooling ratio with low roughness and thickness values. Video examples of the attack can be found here¹. Detailed description of the convergence process regarding this attack can be found in the supplementary material.

4.3.1 Thickness Vs. Roughness

In order to illustrate the trade-off between β_1 and β_2 under single video attacks, we have selected a video sample (kinetics test set) on which we developed two different flickering attacks by solving Equation (1) (separately) under the single video attack settings ($N = 1$). As described in Section 3.4, the β_j 's coefficients control the importance of each regularization term, where β_1 associated with the term that forces the perturbation to be as small as possible in gray-level over the three color channels and β_2 associated with purely temporal terms (norms of the first and second temporal derivatives) forcing the perturbation to be temporally-



Figure 2: Illustration of the trade-off between thickness and roughness in a single video attack as described in Section 4.3.1.

smooth as possible. The first perturbation developed with $\beta_1 = 1$ and $\beta_2 = 0$, minimize the thickness while leaving the roughness unconstrained. The second perturbation developed with $\beta_1 = 0$ and $\beta_2 = 1$, minimize the roughness while leaving the thickness unconstrained. Both of these perturbations cause misclassifications on the I3D model. In order to visualize the difference between these perturbations, we deliberately picked a difficult example to attack which that requires large thickness and roughness. In Figure 2 we plot both attacks in order to visualize the difference between the two cases. Each row combined 8 consecutive frames (out of 90 frames). In the first row, the original (clean) video sample from the “juggling balls” category. In the second row, the adversarial (misclassified) video we developed with $\beta_1 = 1$ and $\beta_2 = 0$ (minimizing thickness). In the third row the adversarial video with $\beta_1 = 0$ and $\beta_2 = 1$ (minimizing roughness). In the fourth row we plot the flickering perturbations with $\beta_1 = 1, \beta_2 = 0$ reaching a thickness of 2.97% and roughness of 4.84%. In the fifth row we plot the flickering perturbations with $\beta_1 = 0, \beta_2 = 1$ reaching a thickness of 7.45% and roughness of 2.20%. As expected, the perturbation with the minimized roughness (last row) is smoother than the one without the temporal constrain (fourth row). Furthermore, even though the thickness of temporal constrained perturbation is much higher (7.45% compare to 2.97%) the adversarial perturbation is less noticeable to the human observer than the one with the smaller thickness. Video examples of the discussed attacks can be found here¹ under “juggling balls”.

4.4. Adversarial Attack Generalization

Unlike single video attack, where the flickering perturbation δ was video-specific, a generalized (or universal) flickering attack is a single perturbation that fools our targeted model with high probability for all videos (from any class

or a specific class). In order to obtain a universal adversarial perturbation across videos we solve the optimization problem in Equation (1) with some attack-specific modifications as described in the following sections.

4.4.1 Class generalization: Untargeted Attack

Adversarial attacks on a single video have limited applicability in the real world. In this section we generalize the attack to cause misclassification to all videos from a specific class with a single generalized adversarial perturbation δ . Our experiments conducted on 100 (randomly picked) out of 400 kinetics classes s.t. for each class (separately) we developed its own δ by solving the optimization problem in Equation (1), where $\{X_n\}_{n=1}^N$ is the relevant class training set split. After developing the class generalization δ we evaluate its fooling ratio performance, thickness and roughness as defined in Section 3.5 on the relevant class evaluation split. The *Single Class* entry in Table 1 shows the statistics of average and standard deviation (across 100 different δ 's) of the fooling ratio, thickness and roughness. Showing that when applying this attack, on average 90.2% of the videos from each class were misclassified. It is obvious that generalization produces perturbation with larger thickness and roughness.

4.4.2 Universal Untargeted Attack

We take one more step toward real world implementation of the flickering attack by devising a single universal perturbation that will attack videos from any class. Constructing such flickering attacks is not trivial due to the small number of trainable parameters ($T \times C$ or 270 in I3D) and in particular that they are independent of image dimensions. Similarly to the previous section, we developed single δ by solving the optimization problem in Equation (1), where $\{X_n\}_{n=1}^N$ is the training set defined as the entire evaluation-split (20K videos) of the Kinetics-400. Once the universal δ was computed, we evaluated its fooling ratio performance, thickness and roughness on a random sub-sample of 5K videos from the kinetics test-split. As can be seen in *Universal Class* entry in Table 1, our universal attack reaches a 93% fooling ratio. One might implement the universal flickering attack as a class-targeted attack using the presented method. In this case, the selected class may affect the efficiency of the adversarial perturbation.

4.5. Time Invariance

Practical implementation of adversarial attacks on video classifiers can not be subjected to prior knowledge regarding the frame numbering or temporal synchronization of the attacked video. In this section we present a time-invariant adversarial attack that can be applied to the recorded scene

without assuming that the perturbation of each frame is applied at the right time. Once this time-invariant attack is projected to the scene in a cyclic manner, regardless of the frame arbitrarily-selected as first, the adversarial pattern would prove to be effective. Similar to the generalized adversarial attacks described in previous subsections, a random shift between the perturbation and the model input was applied during training. The adversarial perturbation in Equation (1) modified by adding the *Roll* operator defined in Equation (6) s.t. $F_\theta(X_n + Roll(\delta, \tau))$ for randomly sampled $\tau \in \{1, 2, \dots, T\}$ in each iteration and on each video during training and evaluation. This time invariance generalization of universal adversarial flickering attack reaches 83% fooling ratio, which is luckily a small price to pay in order to approach real-world implementability.

5. Additional models, baseline comparisons and transferability

In order to demonstrate the effectiveness of the flickering adversarial attack (universal in particular) we applied selected attacks to other relevant models and compared between the proposed universal flickering attack to other baseline attacks (Section 5.2) and validate that our attack is indeed transferable [27] across models (Section 5.3).

5.1. Targeted Models

Similar to the previous experiment we follow the white-box setting. In the following experiments we apply our attack on three different models **MC3**, **R3D**, **R(2+1)D** (pre-trained on the Kinetics Dataset) from [29] which discuss several forms of spatiotemporal convolutions and study their effects on action recognition. All three model are based on 18 layers ResNet architecture [6], accepting spatial and temporal dimensions of: $T = 16, H = 112, W = 112, C = 3$. Implementation details can be found in this paper's supplementary material.

5.2. Baseline comparison

Following the introduction of the first flickering attack against video action recognition models, a baseline comparison of the effectiveness of the universal attack is presented against several types of random flickering perturbations. We developed a universal flickering perturbation δ^F on model F (I3D, R(2+1)D, etc.) with respect to the Kinetics Dataset by solving the optimization problem defined by Equation (1). Following Equation (1) we constrained the ℓ_∞ norm of δ^F by clipping s.t. $\|\delta^F\|_\infty = \max |\delta^F| \leq \zeta$ for some ζ .

In order to evaluate the Fooling ratio of any δ (and in particular δ^F) on some model F we define the evaluation set $\mathbb{X} = \{X_i\}_{i=1}^M$ where $X_i = [x_1^i, x_2^i, \dots, x_T^i]$ is i^{th} evaluation video consisting of T consecutive frames. On top of \mathbb{X} we define the adversarial evaluation set $\hat{\mathbb{X}}_\delta = \{\hat{X}_i\}_{i=1}^M$ where,

$\hat{X}_i = [x_1^i + \delta, x_2^i + \delta, \dots, x_T^i + \delta]$ for all i . Therefore, the fooling ratio is calculated by evaluating F on \hat{X}_δ . In the following experiments we use the same evaluation set \mathbb{X} .

Given a flickering universal adversarial perturbation δ^F developed on model F , we define the following random flickering attacks:

$\delta_U^F \sim \mathcal{U}(\min \delta^F, \max \delta^F)$: Random variable uniformly distributed between the minimal and maximal values of δ^F .

δ_{MinMax}^F : Each element is drawn from the set $\{\min \delta^F, \max \delta^F\}$ with equal probability.

$\delta_{shuffle}^F$: A random shuffle of δ^F along the frames and color channels. Table 2 shows the results of our experiments where each experiment (different ℓ_∞ [%]) was performed as follows:

1. For given ζ we developing δ^F for each one of our four attacked models: I3D, R(2+1)D, R3D and MC3.
2. For each δ^F we developed $\delta_U^F, \delta_{MinMax}^F, \delta_{shuffle}^F$ as described earlier.
3. On each model F we evaluate the fooling ratio of the following perturbation: Random flickering ($\delta_U^F, \delta_{MinMax}^F, \delta_{shuffle}^F$), universal flickering developed upon other models and universal flickering δ^F .

In our experiments the ℓ_∞ [%] norm of δ is represented as the percentage of the allowed pixel intensity range ($V_{max} - V_{min}$). e.g., if $V_{max} = 1, V_{min} = -1$ and ℓ_∞ [%] = 10 than $\zeta = 0.2$. In order to obtain statistical attributes we performed the experiments by re-perturbing the random generated δ 's ($\delta_U^F, \delta_{MinMax}^F, \delta_{shuffle}^F$). As shown in Table 2 we performed the experiments over several values of ℓ_∞ [%]: 5, 10, 15 and 20. The columns (with models names) represent the attacked model, while the rows represent the type of flickering attacks. Random flickering attacks are located at the first 3 rows of each experiment, followed by the universal flickering attack trained upon other models (except I3D)⁵ – marked with (trns). The universal flickering attack (ours) is located at the last row of each experiment. Each cell holds the fooling ratio result (average and standard deviation in the case of random generated perturbations) when evaluating the model on the data with the relevant attack. As can be seen, the universal flickering attack demonstrates superiority across all four models, over the transferable attacks and the random flickering attacks. In addition to Table 2, additional analysis is presented in the supplementary material.

5.3. Transferability across Models

Transferability [27] is defined as the ability of an attack to influence a model which was unknown to the attacker

⁵ The transferability between I3D to the other models (and vice versa) were not evaluated because the input of the models is not compatible.

ℓ_∞ [%]	Attack \ Model	I3D	R(2+1)D	R3D	MC3
5	Random Uniform	8.4± 0.6%	4.9± 0.8%	8.3± 1.8%	11.0± 1.9%
	Random MinMax	12.2± 0.7%	9.0± 2.3%	15.8± 3.5%	17.4± 3.8%
	Flickering shuffle	11.9± 0.6%	9.4± 1.7%	16.4± 3.3%	16.5± 2.5%
	R(2+1)D (trns)	-	-	27.6%	18.4%
	R3D (trns)	-	14.9%	-	24.0%
MC3 (trns)	-	12.3%	31.4%	-	
	Flickering	26.2%	23.3%	34.3%	41.3%
10	Random Uniform	14.2± 1.2%	10.7± 3.3%	20.2± 5.3%	17.9± 3.1%
	Random MinMax	23.6± 2.4%	19.2± 4.8%	36.7± 6.3%	30.0± 3.7%
	Flickering shuffle	22.9± 2.1%	18.3± 5.5%	31.9± 7.2%	25.9± 3.7%
	R(2+1)D (trns)	-	-	52.7%	38.4%
	R3D (trns)	-	30.6%	-	35.6%
MC3 (trns)	-	25.9%	50.5%	-	
	Flickering	58.4%	47.2%	70.4%	55.3%
15	Random Uniform	20.3± 2.1%	16.0± 4.7%	26.2± 4.7%	24.2± 1.8%
	Random MinMax	34.2± 3.1%	28.1± 7.9%	48.6± 7.4%	36.4± 4.9%
	Flickering shuffle	29.3± 3.1%	28.7± 5.0%	44.6± 8.7%	35.3± 2.8%
	R(2+1)D (trns)	-	-	64.4%	48.4%
	R3D (trns)	-	39.5%	-	50.7%
MC3 (trns)	-	40.7%	66.1%	-	
	Flickering	78.1%	62.7%	83.4%	73.3%
20	Random Uniform	32.1± 3.1%	22.2± 5.7%	37.1± 4.0%	30.0± 4.5%
	Random MinMax	48.0± 4.5%	42.0± 3.0%	54.6± 11.0%	44.0± 5.0%
	Flickering shuffle	42.0± 3.6%	39.0± 8.0%	57.6± 6.4%	47.1± 4.7%
	R(2+1)D (trns)	-	-	74.6%	59.2%
	R3D (trns)	-	58.5%	-	60.7%
MC3 (trns)	-	55.8%	70.4%	-	
	Flickering	93.0%	79.0%	90.3%	77.1%

Table 2: Baseline comparison of the universal flickering attack to several types of random flickering attacks and transferability across different models.

when developing the attack. We examined the transferability of the flickering attack on different models of the same input type. As seen in Table 2, for each ℓ_∞ [%] we evaluate the the fooling ratio of attacks that was trained on different models (trns). The high effectiveness of the attack applied across models indicates that our attack is transferable between these different models, e.g., attack that was developed on R(2+1)D with ℓ_∞ [%] = 20 achieved 74.6% fooling ratio when applied on R3D model compared to 90.3%.

6. Over-the-Air Real world demonstration

The main advantage of the flickering attack, unlike the majority of adversarial attacks in published papers, is its real-world implementability. In this section we demonstrate, for the first time, the flickering attack in a real world scenario. We used an RGB led light bulb and controlled it through Wifi connection. Through this connection we were able to control the red, green and blue illumination values separately, and create almost any of the previously developed adversarial RGB patterns introduced in this paper (Figure 1 depicts the modeling of our digital domain attack in the real-world). As in [14, 3], we have applied several constraints for better efficiency of the adversarial manipulations in real-world, such as temporal invariance (Section 4.5) and increased smoothness to address the fi-

nite rise (or fall) time of the RGB bulb (Section 3.4.2). Because the adversarial patterns presented here have positive and negative amplitude perturbations, the baseline illumination of the scenario was set to around half of the possible maximum illumination of the bulb. A chromatic calibration of the RGB intensities was performed in order to mitigate the difference of the RGB illumination of the light bulb and RGB responsivity of the camera, which was obviously not the same and also included channel chromatic crosstalk. The desired scenario for the demonstration of the attack includes a video camera streaming a video filmed in a room with a Wifi-controlled RGB light bulb. A computer sends over Wifi the adversarial RGB pattern to the bulb. A figure performs actions in front of the camera. Implementation and hardware details can be found in the supplementary material. We demonstrate our over-the-air attack in two different ways, scene-based and universal flickering attack.

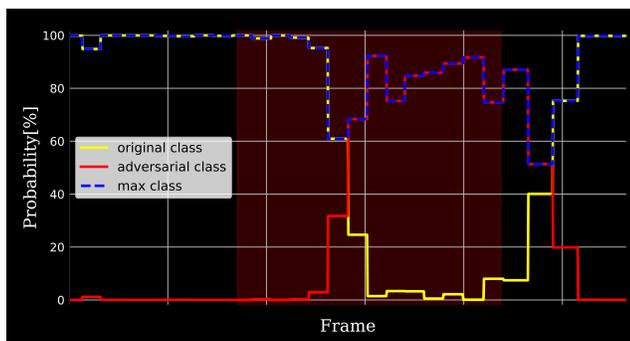


Figure 3: Example of our Over-the-Air scene based attack. The plot was taken from the “ironing” video example².

6.1. Over-the-Air Scene-based Flickering Attack

In this attack, we assume prior knowledge of the scene and the action. Therefore, similar to a single video attack (Section 4.3) we will develop a scene dedicated attack. In this approach we record a clean video (without perturbation) of the scene we would like to attack. For the clean recording we develop a time-invariant digital attack as described in the paper. Once we have the digital attack, we transmit it to a “similar” scene (as described in the supplementary material) in an over-the-air approach. Video examples of our scene based over-the-air adversarial attack can be found here². Figure 3 shows the probability of a real example of our scene based over-the-air attack of the “ironing” action, where the x-axis (Frame) represents prediction time step and the y-axis (Probability) represents the output probability of the I3D model for several selected classes. The area shaded in red represents the period of time the scene was attacked. As described in the legend, the yellow graph is the true class (“ironing”) probability, the red graph is the adversarial class (“drawing”) probab-

ity and the dashed blue graph represents the probability of the most probable class the classifier predicts each frame. It can be seen that when the scene is not attacked (outside the red area) the model predicts correctly the action being performed (dashed blue and yellow graphs overlap). Once the scene is attacked, the true class is suppressed and the adversarial class is amplified. At the beginning (end) of the attack, it can be seen that there is a delay from the moment the attack begins (ends) until the model responds to the change due to the time required (90 frames) to fill the classifier’s frame buffer and perform the prediction.

6.2. Over-the-Air Universal Flickering Attack

This section deals with the case where we do not have any prior knowledge regarding the scene and action we wish to attack. Therefore, we would like to develop a universal attack that will generalize to any scene or action. In this approach, we will use a universal time-invariant attack as described in the paper. Once we have the digital attack, we transmit it to the scene in an over-the-air approach. Video examples of our universal over-the-air attack can be found here³. Since our approach is real-world applicable, and thus we require universality and time-invariability perturbation (no need to synchronize the video with the transmitted perturbation), the pattern is visible to the human observer.

7. Conclusions and future work

The flickering adversarial attack was presented, for the first time, for several models and scenarios summarized in Tables 1, 2. Furthermore, this attack was demonstrated in the real world for the first time. The flickering attack has several benefits, such as the relative imperceptibility to the human observer in some cases, achieved by small and smooth perturbations as can be seen in the videos we have posted¹. The flickering attack was generalized to be universal, demonstrating superiority over random flickering attacks on several models. In addition, the flickering attack has demonstrated the ability to transfer between different models. The flickering adversarial attack is probably the most applicable real-world attack amongst any video adversarial perturbation this far, as was shown^{3,2}. Thanks to the simplicity and uniformity of the perturbation across the frame which can be achieved by subtle lighting changes to the scene by illumination changes. All of these properties make this attack implementable in real-world scenarios.

In extreme cases where generalization causes the pattern to be thick enough to be noticed by human observers, the usage of such perturbations can be relevant for non-man-in-the-loop systems or cases where the human observer will see image-flickering without associating this flickering with an adversarial attack. In the future, we may expand the current approach to develop defensive mechanisms against adversarial attacks of video classifier networks.

References

- [1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 3
- [2] 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. 1, 2, 4
- [3] Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. 7
- [4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 1
- [5] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1, 3
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [7] Eddy Ilg, Nikolaus Mayer, Tommo Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 3
- [8] Nathan Inkawhich, Matthew Inkawhich, Yiran Chen, and Hai Li. Adversarial attacks for optical flow-based action recognition classifiers. *arXiv preprint arXiv:1811.11875*, 2018. 1, 3
- [9] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35:221–231, 2013. 2
- [10] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. Black-box adversarial attacks on video recognition models. *ArXiv*, abs/1911.09449, 2019. 1, 3
- [11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2, 5
- [12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 1
- [13] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 3
- [14] Juncheng Li, Frank Schmidt, and Zico Kolter. Adversarial camera stickers: A physical camera-based attack on deep learning systems. In *International Conference on Machine Learning*, pages 3896–3904. PMLR, 2019. 7
- [15] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V. Krishnamurthy, Amit K. Roy-Chowdhury, and Ananthram Swami. Stealthy adversarial perturbations against real-time video classification systems. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*, 2019. 3
- [16] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60 – 88, 2017. 1
- [17] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 1
- [18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 1
- [19] Roberto Rey-de Castro and Herschel Rabitz. Targeted non-linear adversarial perturbations in images and videos. *arXiv preprint arXiv:1809.00958*, 2018. 3
- [20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2
- [21] Hasim Sak, Andrew W. Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 338–342, 2014. 2
- [22] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017. 1
- [23] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 2
- [24] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1
- [25] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neu-*

- ral Information Processing Systems 27*, pages 1988–1996. Curran Associates, Inc., 2014. [1](#)
- [26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. [2](#)
- [27] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. [1](#), [6](#), [7](#)
- [28] 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](#)
- [29] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. [2](#), [6](#)
- [30] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517, 2017. [2](#)
- [31] Lei Wang, Yangyang Xu, Jun Cheng, Haiying Xia, Jianqin Yin, and Jiaji Wu. Human action recognition by learning spatio-temporal features with deep neural networks. *IEEE Access*, 6:17913–17922, 2018. [2](#)
- [32] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2017. [1](#)
- [33] Xingxing Wei, Jun Zhu, Sha Yuan, and Hang Su. Sparse adversarial perturbations for videos. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8973–8980, 2019. [1](#), [3](#)
- [34] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang. Heuristic black-box adversarial attacks on video recognition models. In *AAAI*, pages 12338–12345, 2020. [1](#), [3](#)