Merry Go Round: Rotate a Frame and Fool a DNN

Daksh Thapar, Aditya Nigam
Indian Institute of Technology Mandi
dakshthapar.github.io, faculty.iitmandi.ac.in/~aditya

Chetan Arora
Indian Institute of Technology Delhi
www.cse.iitd.ac.in/~chetan

Abstract

A large proportion of videos captured today are first person videos shot from wearable cameras. Similar to other computer vision tasks, Deep Neural Networks (DNNs) are the workhorse for most state-of-the-art (SOTA) egocentric vision techniques. On the other hand DNNs are known to be susceptible to Adversarial Attacks (AAs) which add imperceptible noise to the input. Both black-box, as well as white-box attacks on image as well as video analysis tasks have been shown. We observe that most AA techniques basically add intensity perturbation to an image. Even for videos, the same process is essentially repeated for each frame independently. We note that definition of imperceptibility used for images may not be applicable for videos, where a small intensity change happening randomly in two consecutive frames may still be perceptible. In this paper we make a key novel suggestion to use perturbation in optical flow to carry out AAs on a video analysis system. Such perturbation is especially useful for egocentric videos, because there is lot of shake in the egocentric videos anyways, and adding a little more, keeps it highly imperceptible. In general our idea can be seen as adding structured, parametric noise as the adversarial perturbation. Our implementation of the idea by adding 3D rotations to the frames, reveal that using our technique, one can mount a black-box AA on an egocentric activity detection system in one-third of the queries compared to the SOTA AA technique.

1. Introduction

Despite achieving superior performance on a variety of computer vision tasks [3, 11, 12, 33], Deep Neural Networks (DNNs) remain remarkably susceptible to imperceptible adversarial perturbations [37]. The goal of an adversarial attack (AA) is, given a clean image, I, create an adversarial perturbation (P), which when added to the clean image, generates an adversarial sample $I_{\text{adv}} = I + P$, which tricks a DNN model into producing an incorrect prediction. Since the purpose is to attack a system, the perturbation should be imperceptible to the humans.

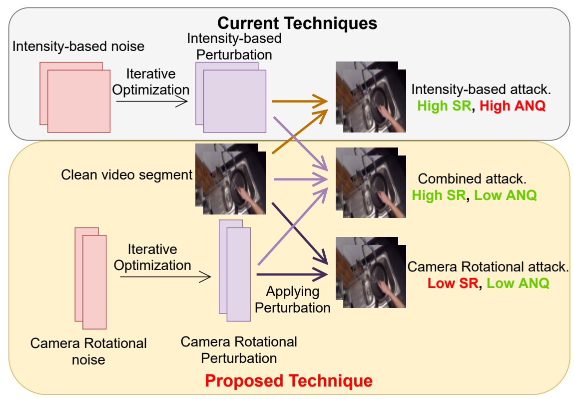

Figure 1. A brief pipeline for the proposed system. SR denote Success Rate, and ANQ denotes Average Number of Queries. Successful Attack demands high SR and low ANQ. For a given input video of size $T \times H \times W \times C$, where T is the number of frames, and H, W, and C are height, width, and channel respectively of each frame, an intensity based attack needs to predict $T \times H \times W \times C$ parameters. Whereas, our proposed parametric perturbation attack, using rotation based transformation, predicts only $T \times 3$ parameters. This reduces the query budget to predict the parameters. Geometric transformations are natural perturbations and do not disturb semantic integrity of an image or a video.

The simplest setting to mount such an AA is when the adversary gets full access to the model (M), including input(X)/output(Y), and the exact gradients (G). One can, then, simply backpropagate the loss corresponding to the desired (incorrect) output, and use it guide the perturbation in the input [16, 25, 37]. The setting is called *white box* attacks, but is usually impractical in real life, due to unavailability of the full access to the model. The alternate setting is the *black box* setting when an adversary has access to X, and Y, but not G. In this formulation the primary challenge becomes estimating the gradient at the input without having access to G [6, 16, 17]. The quality of an AA technique is usually determined by how imperceptible the P is, and additionally in case of black box attacks, how many (X,Y) pairs a technique needs to find a P corresponding to a particular I.

Researchers have shown both white box and black box attacks for a variety of DNN models across range of tasks [37]. Further, relevant to our context, the attacks have been shown when the input to the model is an image [16,17], or a video [20,49]. Our focus in this paper is on mounting black box adversarial attacks on video analysis (VA) systems .

We note that most of the techniques for AA on a VA system trivially extends the black box pipeline from images to videos. The videos are broken down into frames, and adversarial examples are created by adding random perturbations in the pixel intensities [20, 49]. For a successful attack, these methods require a large number of queries on the target model. For example [20] requires 23K queries on an average for generating a single adversarial sample. We would like to emphasize that a frame-wise attack, using intensity-based noise, do not coordinate the adversarial perturbations between consecutive frames. While a change in intensity level of a few individual pixels may be imperceptible in an individual frame, when played as a video, such random flashes are easily detected by a human being.

One of the key ideas of this paper is to parameterize the perturbation. The parameterization has two advantages, (1) it is easier to regularize within, and across the frames, and (2) one can perturb a large number of pixels, by estimating only a few parameters, thus reducing the query budget, an important consideration in a black box attack. While, the idea of parametric perturbation is generic and can be used in variety of settings, given our focus to videos, we consider it for attack on VA systems, and even more specifically, on egocentric VA systems.

We observe that one of the simplest ways to perform coordinated change in intensity levels of large number of pixels, across frames of a video, is by geometrically transforming each frame. The transformation will cause change in the optical flow, which is an important cue for many VA tasks. At the same time, performing frame-wise geometric transformation maintains semantic integrity of frame contents, keeping it imperceptible to human beings.

Contributions: The key contributions of this work are:

1. We propose to add novel parametric perturbations to mount an AA attack against a computer vision system.
2. For a VA system, we suggest use of geometric transformations to implement such parametric perturbations.
3. We propose a novel DNN architecture for predicting a mix of intensity, and geometric perturbations which can successfully fool a VA system to carry out black box AA attack.
4. Our exhaustive experiments on multitude of benchmark datasets, and VA tasks for egocentric, and third person videos show that our proposed architecture outperforms SOTA techniques, managing to fool a DNN in one-third of the queries as needed by the SOTA.

2. Related Work

Adversarial Attacks: Szegedy *et al.* [37] have shown that by computing a small noise on the original image, one can create an adversarial example. Papernot *et al.* [25] have shown that a black box attack can be carried out on a target model by transferring the adversarial examples of a local trained network. However, such a technique still requires knowledge of the dataset and training procedure of the target model. Natural Evolutionary Strategies have been extended in [16] to perform gradient estimation. Ilyas *et al.* [17] have shown that time and data-dependent priors can reduce the number of queries in black box attacks. The meta-based method has been proposed by Du *et al.* [6] for black box attacks on image analysis models. However, little work has been done on attacking DNNs for VA. Further, to the best of our knowledge, there is no AA proposed for egocentric VA models.

Adversarial Attacks on Video Analysis Models: For third-person videos, Wei *et al.* [44] have investigated the sparsity and propagation of adversarial perturbations across videos for creating a white-box attack. Li *et al.* [22] have proposed Generative Adversarial Networks to synthesize adversarial examples for a video classification DNN. Inkawhich *et al.* [18] have proposed an FGSM [10] style of attacks for attacking a two-stream video classifier. Chen *et al.* [4] added a few fake frames to attack video classification DNNs. The first black-box video attack is proposed by Jiang *et al.* [20], where they have used an ImageNet pre-trained model to create a gradient for each video frame and refined them by using natural-evolution-strategies [16]. More recently, [45, 47] perturb only a few selected frames rather than the whole video. In [49] a motion based sampler for perturbing every frame in the video has been proposed.

Third-person Video Analysis: Recent methods for third-person video action recognition utilize 3D CNNs [2, 7, 19, 40, 43, 50]. 3D CNNs extend 2D filters in temporal dimensions to extract spatio-temporal features from videos. Since early 3D models [19, 40] are hard to train, many follow-up works have been proposed [2, 7, 31, 41]. Two-stream methods proposed in [34] combine a spatial network using RGB images and a temporal network taking optical flow input. Optical flow information has also been found beneficial in few-shot video classification [51].

First-person Video Analysis: Some notable works in general egocentric video analysis include camera wearer's activity and action recognition [1, 21, 28–30, 35, 36, 42], wearer gaze estimation [15], temporal segmentation [24], and video summarization [32, 46]. Another uniquely egocentric video task is recognizing the wearer capturing the video. The task has attracted lot of attention in recent years [7, 8, 13, 14, 23, 26, 27, 30, 38, 39].

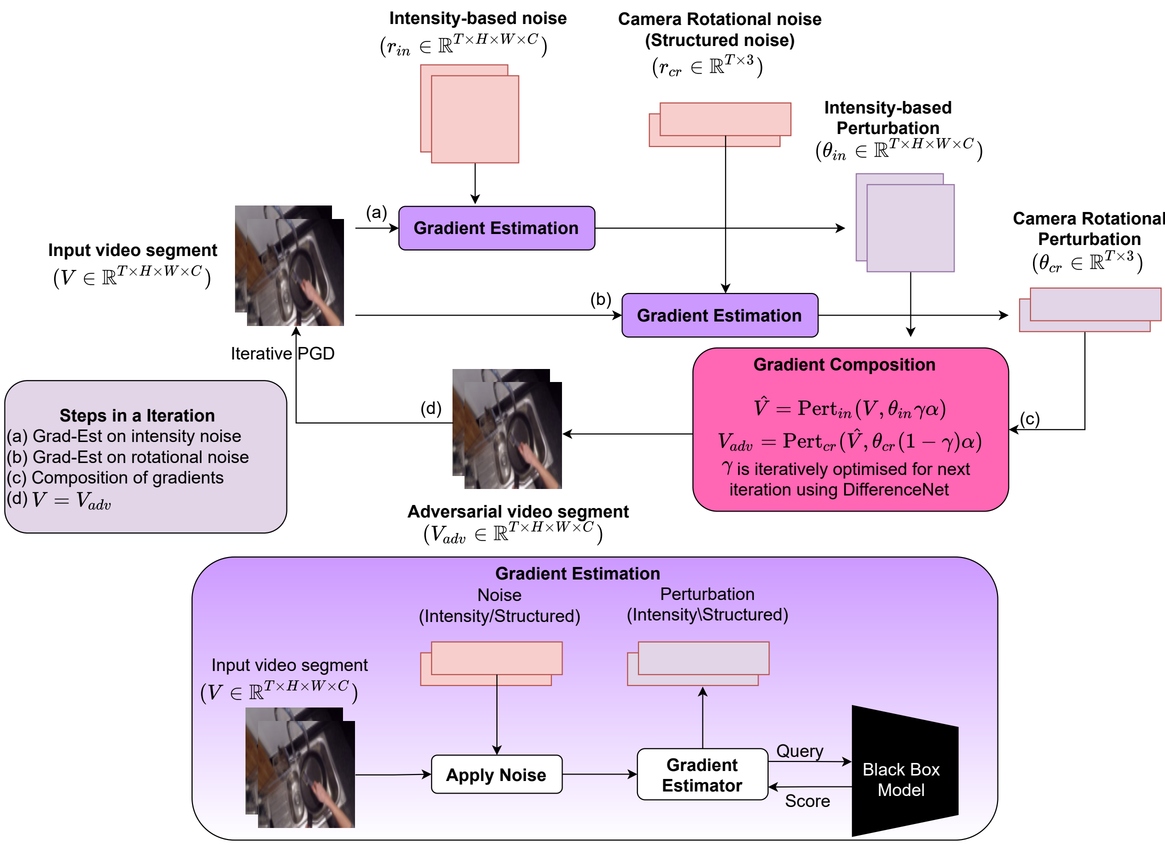

Figure 2. Overview of our framework for black-box video attack. The steps in the iterations are numbered.i) Compute Gradient estimation for intensity-based noise; ii) Compute Gradient estimation for camera rotational noise; iii) Composing the gradients utilizing DifferenceNet (Extracts semantic difference between $\boldsymbol{I}$ and $\boldsymbol{I_{adv}}$); iv) Use the estimated gradient to perform iterative projected gradient descent (PGD) optimization on the video.

3. Proposed Methodology

3.1. Gradient Estimation

We consider a DNN model f, which has been pre-trained for some VA task. The model takes as input a video $V \in \mathbb{R}^{T\times H\times W\times C}$, where T, H, W, and C represent video length, height, width, and number of channels (in each frame) respectively. Assuming a video classification model, the output of f is a label $y \in \{1, \dots K\}$, where K is the number of classes. The goal of an adversarial attack is, given an input video V, generate an adversarial video V_{adv} which minimises the loss function:

$$\mathcal{L} = \max(\mathbf{l}_y - \max_{k\neq y}(\mathbf{l}_k), 0). \tag{1}$$

Here $\mathbf{l}$ is the *logit* vector corresponding to input V_{adv}, and $\mathbf{l}_i$ is the value of i^{th} element (corresponding to class i) of the vector. Minimizing $\mathcal{L}$ confuses the model with the second most confident class prediction for the untargeted adversarial attack. For the targeted attack $\max_{k\neq y}(\mathbf{l}_k)$ can be replaced by the logit of the corresponding class. To simplify the notation, in the rest of the paper we simply use $\mathcal{L}(V, y)$ instead of $\mathcal{L}(f(V), y)$. The adversarial video V_{adv} is chosen as:

$$\begin{aligned} &\arg\min_{V_{\text{adv}}} \mathcal{L}(f(V_{\text{adv}}), y) \\ &\text{s.t.} \quad \text{dist}(V_{\text{adv}}, V) \leq \text{max\_dist}, \\ &\text{and} \quad \#\text{queries} \leq Q. \end{aligned} \tag{2}$$

We can model V_{adv} using any perturbation parameterized by $\theta \in \mathbb{R}^{T\times d}$, where d is the dimension of θ, s.t. $V_{\text{adv}} = \text{Pert}(V, \theta)$. Here, the function $\text{Pert}(V, \theta)$ applies the perturbation parameterized by θ on the video V. The function Pert will be dependent upon the type of perturbation and is defined in detail in Sec. 3.2. To generate an adversarial video V_{adv}, we need to find an optimal perturbation θ^* s.t.:

$$\begin{aligned} \theta^* &= \arg\min_{\theta} \mathcal{L}(\text{Pert}(V, \theta), y) \\ &\text{s.t.} \quad \|\theta\|_2 \leq k, \\ &\text{and} \quad \#\text{queries} \leq Q. \end{aligned} \tag{3}$$

Here k is the maximum perturbation allowed. We have used ℓ_2 norm for constraining the θ, but any other suitable constraint on the θ could have been used. The above perturbation framework allows us to generalize the adversarial attacks additive, multiplicative, or even some complex non-differentiable perturbations. Moreover, it allows us to design a parametric perturbation of a very low dimension d which is easier to compute in limited query budget.

The key challenge in black-box adversarial attacks is to estimate the gradient of a model. It is because for this setting, the model is not accessible (beyond input, output), and the gradient $\nabla_\theta \mathcal{L}(\text{Pert}(V, \theta), y)$, required for generating V_{adv} cannot be directly computed. Hence, we adopt an iterative optimization strategy suggested in [49] for estimating $\nabla_\theta \mathcal{L}(\text{Pert}(V, \theta), y)$.

It is important to note that for an iterative optimization, we are only interested in the direction of $\nabla_\theta \mathcal{L}(\text{Pert}(V, \theta), y)$ rather than its exact value which also includes the magnitude. Hence, we learn a vector $g \in \mathbb{R}^{T \times d}$ whose direction ($\frac{g}{\|g\|}$) aligns with $\nabla_\theta \mathcal{L}(\text{Pert}(V, \theta), y)$. In order to estimate such a g, we use the following loss function [17]:

$$l(g) = -\langle \nabla_\theta \mathcal{L}(\text{Pert}(V, \theta), y), \frac{g}{\|g\|} \rangle, \tag{4}$$

which is the inverse of directional derivative of $\mathcal{L}$, in the direction of the vector g. The inverse direction of directional derivative provides the direction of g's movement to optimize $l(g)$ and get closer to the desired gradient $\nabla_\theta \mathcal{L}(\text{Pert}(V, \theta), y)$ as:

$$g^* = \arg\min_g (l(g)). \tag{5}$$

In order to compute g^*, we compute the gradient $\nabla_g l(g)$, denoted as Δ. We perform a two-query estimation to the expectation and apply the authentic sampling [17] to get:

$$\Delta = \left[\frac{l(g + \delta r) - l(g - \delta r)}{\delta} \right] r, \tag{6}$$

where $r \in \mathbb{R}^{T \times d} \in \mathcal{N}(0, \frac{1}{d} I)$ is the Gaussian noise, and δ is a small number scaling the magnitude of loss variation. In two-query estimation, r vector acts as a directional candidate for the update of g. We query in the direction of r and in its opposite direction. This gives us a scalar indicating of how good the candidate r is. We scale r accordingly to form our update of g.

Finally, Eq. (4) can be approximated as [17]:

$$\begin{aligned} l(g) &= -\langle \nabla_\theta \mathcal{L}(\text{Pert}(V, \theta), y), \frac{g}{||g||} \rangle \\ &\approx -\frac{\mathcal{L}(\text{Pert}(V, \theta + \epsilon g), y) - \mathcal{L}(\text{Pert}(V, \theta), y)}{\epsilon}, \end{aligned} \tag{7}$$

where ϵ is a small approximation constant. Substituting Eq. (7) into Eq. (6), we get $\text{GE}(V, y, \theta, g)$ as:

$$\begin{aligned} \Delta &= \text{GE}(V, y, \theta, g) \\ &= \left[\frac{\mathcal{L}(\text{Pert}(V, \theta + \epsilon g^+), y) - \mathcal{L}(\text{Pert}(V, \theta + \epsilon g^-), y)}{\epsilon \delta} \right] r, \end{aligned} \tag{8}$$

where $g^+ = g + \delta r$ and $g^- = g - \delta r$.

3.2. Parametric Noise

It can be observed from Eq. (8), that in order to estimate the gradient, we have utilized a random noise (r). For intensity-based noise, $r_{\text{in}} \in \mathbb{R}^{T \times H \times W \times C}$ is used for estimating the gradient $g_{\text{in}} \in \mathbb{R}^{T \times H \times W \times C}$ [49]. This requires one to estimate $T \times H \times W \times C$ parameters for the adversarial attack, which may lead to a high number of queries [49], making such attacks unrealistic in practice.

To overcome these limitations, we have proposed a parametric noise (camera rotational noise r_{cr}) which can suitably alter the geometrical properties of a video for an attack. Since, rotation of the camera can be represented as a 3D vector in Euler space, the proposed noise $r_{\text{cr}} \in \mathbb{R}^{T \times 3}$, requires only $T \times 3$ parameters to be predicted for an adversarial attack. This significantly reduces the number of queries required to predict it in comparison to an intensity-based noise.

We estimate the camera rotational gradient $g_{cr} \in \mathbb{R}^{T \times 3}$ from r_{cr} using gradient estimation, as discussed in the previous section. This allows us to find a new perturbation vector θ, with $\theta_i \in \mathbb{R}^3$ for each frame. Recall, that θ_i corresponds to a 3D rotation for the frame. We compute an Homography using the 3D rotation as $\mathcal{H}_i = K \cdot \theta_i K^{-1}$, where K is the camera internal matrix (assumed identity in our case). The perturbation can be applied on the video as:

$$\text{Pert}_{cr}(V, \theta) = \forall_i (\mathcal{H}_i * V_i), \tag{9}$$

where, V_i is the i^{th} frame in the video V and $*$ denotes the geometric transformation of each frame using the Homography $\mathcal{H}_i$. To ensure that the perturbations are small, we have clipped the magnitude of r_{cr} to 0.18 radians.

We observe that in our experiments the number of queries required to render a successful black-box attack gets substantially reduced by using parametric noise, but at the expense of success rate (refer Sec. 4.2). Hence, we propose to mix it with intensity based perturbation, using a learnable composition parameter, as described in the next section.

3.3. Gradient Composition

In order to address the issue of low success rate using parametric noise, we propose a novel learnable gradient composition framework which suitably combines intensity-based, and parametric perturbations. Such fusion exploits spatio-temporal properties of a particular segment in a video to dynamically adjust the weights of two kinds of perturbation, and achieve lower queries. For example, if there is very small motion between two frames, intensity based noise can be more effective. However, in the case of large temporal movements of objects or camera, the rotational noise can be useful. We propose a Siamese network based architecture , named DifferenceNet, to predict the weight of each perturbation for a frame.

DifferenceNet: The proposed DifferenceNet model is a 3D CNN model (with I3D [2] pipeline) trained to calculate semantic difference between input video (V) and adversarial video (V_{adv}). The task of DifferenceNet is to provide a low difference score to videos which are semantically similar otherwise a high score. This is achieved by training the network with a dual margin contrastive loss function [48]. The network is trained over positive pairs which have the camera rotations between the frames corresponding to actual videos and negative pairs having abrupt rotations between the frames. To create positive and negative pairs, real Homographies ($\mathcal{H}_{\text{real}}$), between the frames from the given dataset D and random/fake Homographies have been generated. Application of $\mathcal{H}_{\text{real}}, \mathcal{H}_{\text{rand}}$ on a video segment V, gives us (V^p, V^n) constituting a positive and negative pair as $(\langle V, V^p \rangle)$, $(\langle V, V^n \rangle)$ respectively. Finally, the trained network is utilized for gradient composition as described below.

Gradient Composition: For a given input V, intensity based perturbation, and camera based perturbations are combined as:

$$\begin{aligned} \hat{V} &= \text{Pert}_{\text{in}}(V, \alpha\gamma\theta_{\text{in}}) \\ V_{\text{adv}} &= \text{Pert}_{\text{cr}}(\hat{V}, \alpha(1-\gamma)\theta_{\text{cr}}), \end{aligned} \tag{10}$$

where, $\gamma \in [0,1]^{T\times 1}$ is the composition parameter and α is a small constant. Since γ depends on semantic difference between (V, V_{adv}), we have utilized DifferenceNet to predict its value:

$$\begin{aligned} d &= \text{DifferenceNet}(V, V_{\text{adv}}) \\ \gamma &= \gamma - \sigma\left(\frac{\delta d}{\delta \gamma}\right), \end{aligned} \tag{11}$$

where σ is a small constant.

3.4. Projected Gradient Descent

Finally, projection gradient descent (PGD) has been utilized to translate gradient estimation and its combination into an efficient Adversarial Example Optimization (AEO). We update intensity based perturbation ($\text{Pert}(V, \theta_{\text{in}})$), camera rotational perturbation ($\text{Pert}(V, \theta_{\text{cr}})$), and composition parameter (γ) in every iteration of PGD. The complete procedure is shown in Algorithm 1.

Algorithm 1: Adversarial Example Optimization (AEO)

```
Input: Original video V, its label y, learning rate α
       for updating adversarial video.
Initialise g_in = 0, g_cr = 0, θ_in = 0, θ_cr = 0 and
  γ = 0.5
while arg max [f(V)] = y do
    Δ_in = GE(V, y, θ_in, g_in)              // Eq 8
    Δ_cr = GE(V, y, θ_cr, g_cr)              // Eq 8
    g_in = g_in − ηΔ_in           // Grad. Update
    g_cr = g_cr − ηΔ_cr           // Grad. Update
    θ_in = θ_in − g_in            // Param. Update
    θ_cr = θ_cr − g_cr            // Param. Update
    V̂ = Pert_in(V, θ_in γα)                  // Grad.
      Composition
    V_adv = Pert_cr(V̂, θ_cr(1 − γ)α)         // Grad.
      Composition
    d = DifferenceNet(V, V_adv)
    γ = γ − α × δd/δγ
    V = V_adv
Output: V_adv
```

4. Experiments and Results

In this section, we provide the details of the experimental analysis performed to validate the efficacy of the proposed method. We start with the details of the experimental setup, including details about the datasets used, target DNN models attacked, attack setting, and evaluation metrics. Finally, we show the comparative analysis and ablation study using both quantitative and qualitative experiments.

4.1. Dataset and Evaluation

Datasets: We perform video attacks on three video tasks: third-person action recognition using Kinetics-400 [2] dataset, first-person activity recognition via Epic-Kitchens [5] dataset, and first-person wearer recognition using IITMD-WFP [38] dataset. Kinetics-400 is a large-scale dataset that has around 300K videos in 400 classes. Epic-Kitchens is a first-person activity recognition dataset that consists of 55 hours of egocentric videos from 32 subjects and contains 125 labeled activities performed by the subjects. IITMD-WFP dataset [38] consists of 3.1 hours of videos captured from 31 different subjects. The dataset has been captured under indoor and outdoor scenarios.

DNN Video Analysis Models Used for Experiments: For third-person video action recognition, we follow the experimental setup of [49]. We choose video action recognition model I3D [2] as our black-box model. For I3D training on Kinetics-400, we train it from ImageNet initialized weights. For first-person activity recognition, we choose

Dataset	Method	ANQ	SR%
Kinetics-400	V-Bad [20]	4,047	99.75
	ME-Sampler [49]	2,717	99.00
	Proposed	1,257	99.33
Epic-Kitchens	V-Bad [20]	8,483	99.71
	ME-Sampler [49]	7,326	100.00
	Proposed	3,564	100.00
IITMD-FPR	V-Bad [20]	5,480	94.67
	ME-Sampler [49]	6,025	92.62
	Proposed	3,487	96.33

Table 1. Untargeted attacks on Kinetics-400, Epic-Kitchens, and IITMD-FPR. The attacked models are I3D, Rolling-Unrolling LSTM, and EgoGaitNet respectively.

Dataset	Method	ANQ	SR%
Kinetics-400	V-Bad [20]	23,182	92.95
	ME-Sampler [49]	11,120	94.67
	Proposed	6,234	95.82
Epic-Kitchens	V-Bad [20]	44,326	84.23
	ME-Sampler [49]	22,541	89.12
	Proposed	15,283	91.56
IITMD-FPR	V-Bad [20]	34,382	82.19
	ME-Sampler [49]	18,759	86.67
	Proposed	9,910	87.33

Table 2. Targeted attacks on Kinetics-400, Epic-Kitchens, and IITMD-FPR. The attacked models are I3D, Rolling-Unrolling LSTM, and EgoGaitNet respectively.

Dataset	Method	ANQ	SR%
Kinetics-400	Only Intensity	3,569	99.0
	Only Rotation	1,067	38.19
	Manual Composition	1,884	62.50
	Proposed	1,257	99.33
Epic-Kitchens	Only Intensity	8,238	100.00
	Only Rotation	3,286	62.81
	Manual Composition	4,467	79.67
	Proposed	3,564	100.00
IITMD-FPR	Only Intensity	6,356	95.23
	Only Rotation	3,286	58.42
	Manual Composition	4,019	72.48
	Proposed	3,487	96.33

Table 3. Ablation study on Kinetics-400, Epic-Kitchens, and IITMD-FPR. The attacked models are I3D, Rolling-Unrolling LSTM, and EgoGaitNet respectively.

Rolling-Unrolling LSTM [9] as our black-box model. The pre-trained weights of the model have been provided by the authors. For first-person wearer recognition, we choose EgoGaitNet [38] model. We perform the training procedure as suggested by the authors, and using the code provided.

Attack Setting [49]: We perform both untargeted and targeted attacks under limited queries. An untargeted attack requires the given video to be mis-classified to any wrong label, whereas a targeted attack requires classifying it to a specific label. We randomly select one video from each category for each dataset following the setting in [49]. The target model correctly classifies all selected original videos. We normalize the pixels between 0-1. We constrain the maximum intensity perturbation to 0.03, maximum camera rotational perturbation to 0.18 radians, and maximum queries to Q = 60,000 for untargeted attack. For targeted attack we choose maximum intensity perturbation to 0.05, maximum camera rotational perturbation to 0.18 radians, and maximal queries to Q = 200,000. If a technique is not able to find adversarial perturbation within these constraints, we record it as having consumed Q queries.

Evaluation Metric [49]: We use the average number of queries (ANQ) required in generating adversarial examples and the attack success rate (SR) as the metrics for comparison. ANQ measures the average number of queries made in attacking across all videos, and SR gives the overall success rate in attacking within a query budget Q. Thus, a smaller ANQ and higher SR are desirable.

4.2. Quantitative Comparison

Untargeted Attacks: We report the effectiveness of our proposed method compared to SOTA in Tab. 1. We compare with V-BAD [20], and ME-Sampler [49]. To the best of our knowledge these are the only two video based adversarial attack models with the source code available. We see that our technique achieves comparable SR as the SOTA, while taking a fraction of query budget in comparison. We also report the comparative performance on top-5 performing classes of each of the attacked model in Tab. 4.

Targeted Attack: We report the results of the targeted attacks in Tab. 2. We also report the results of top-5 performing classes of each attacked model in Tab. 5. Similar to untargeted attacks, here also we observe similar SR performance and a large improvement in query budget. For example, on Epic-Kitchens, our method consumes only 15,283 queries, in comparison to 44,326 by V-BAD and 22,541 by ME-Sampler, an improvement of almost 3×. Even for Kinetics dataset, we outperform V-BAD and ME-Sampler by saving 16,948 and 4,886 queries, respectively, and achieve a comparable success rate.

Dataset	Method	Class 1		Class 2		Class 3		Class 4	
		ANQ	**SR%**	**ANQ**	**SR%**	**ANQ**	**SR%**	**ANQ**	**SR%**
Kinetics-400	V-Bad [20]	4,618	99.54	4,975	99.57	4,857	99.83	4,573	99.85
	ME-Sampler [49]	2,246	99.32	2,554	98.71	2,794	98.68	2,825	99.46
	Proposed	1,851	99.35	1,719	99.40	1,548	99.31	1,881	99.24
Epic-Kitchens	V-Bad [20]	8,421	99.61	8,156	99.72	8,195	99.70	8,711	99.86
	ME-Sampler [49]	7,672	100.00	7,914	100.00	7,574	100.00	7,057	100.00
	Proposed	6,496	100.00	6,944	100.00	6,700	100.00	6,994	100.00
IITMD-FPR	V-Bad [20]	5,836	94.11	5,706	94.51	5,517	93.73	5,225	93.57
	ME-Sampler [49]	5,720	92.53	5,661	91.34	6,566	91.77	5,970	91.06
	Proposed	3,531	95.97	3,718	96.38	3,304	96.32	3,087	96.20

Table 4. Untargeted attacks on top-4 performing classes of Kinetics-400, Epic-Kitchens, and IITMD-FPR. The attacked models are I3D, Rolling-Unrolling LSTM, and EgoGaitNet respectively.

Dataset	Method	Class 1		Class 2		Class 3		Class 4	
		ANQ	**SR%**	**ANQ**	**SR%**	**ANQ**	**SR%**	**ANQ**	**SR%**
Kinetics-400	V-Bad [20]	23,059	91.74	27,234	93.14	25,735	93.47	20,838	93.15
	ME-Sampler [49]	11,217	95.24	11,181	94.62	11,329	95.27	10,959	93.51
	Proposed	6,414	95.87	6,037	95.65	6,163	96.03	5,894	95.93
Epic-Kitchens	V-Bad [20]	43,646	82.55	43,436	83.15	46,424	84.92	48,762	85.91
	ME-Sampler [49]	22,040	87.96	22,159	88.94	22,494	89.90	22,820	89.65
	Proposed	15,037	92.34	15,071	91.49	15,118	91.91	14,988	90.21
IITMD-FPR	V-Bad [20]	30,338	82.01	35,508	81.36	31,781	81.67	34,590	81.95
	ME-Sampler [49]	18,553	86.26	18,888	87.11	18,269	87.44	18,493	85.94
	Proposed	9,908	87.23	10,471	87.81	10,337	87.92	9,902	86.72

Table 5. Targeted attacks on top-4 performing classes of Kinetics-400, Epic-Kitchens, and IITMD-FPR. The attacked models are I3D, Rolling-Unrolling LSTM, and EgoGaitNet respectively.

4.3. Qualitative Analysis

The comparative qualitative analysis of the proposed framework with ME-Sampler [49] is shown in Fig. 3. We have shown the analysis for three video segments, choosing the middle frame from each video. For detailed analysis, please refer to the supplementary material. The first column shows the original frame, the second column shows the attacked frame using ME-Sampler [49], and the third column shows the attacked frame using our proposed technique. We have also mentioned the number of queries required for the successful attack for each frame. It is evident from the figure that our proposed framework, similar to ME-Sampler, produces imperceptible perturbation to the video frame. However, our proposed framework requires substantially smaller number of queries for successful attack.

4.4. Ablation Study

Intensity based Vs Geoemtric Perturbation: We have conducted ablation study to understand importance of various components of the proposed architecture. Our method introduces a mix of intensity based and geometric noise. In Tab. 3 we show the results, when only one of the noise type is used for perturbation. We see that only intensity based attack causes much more query to generate the perturbation, whereas rotation based attacks require much lesser queries but also a much lower success rate. Combining the both as in the proposed framework, achieves high success rate at a lower query budget.

Manual Vs Learnt γ: The composition factor to combine the intensity based and geometric perturbation is automatically learnt by our model using DifferenceNet. In Tab. 3 we also show the results after setting composition weight manually. One can see that similar to geometric perturbation, the configuration achieves low success rate, at a low query budget. Automated learning of composition weight gives best results, thus validating the need of DifferenceNet.

Distribution of γ: One of the key components of our model

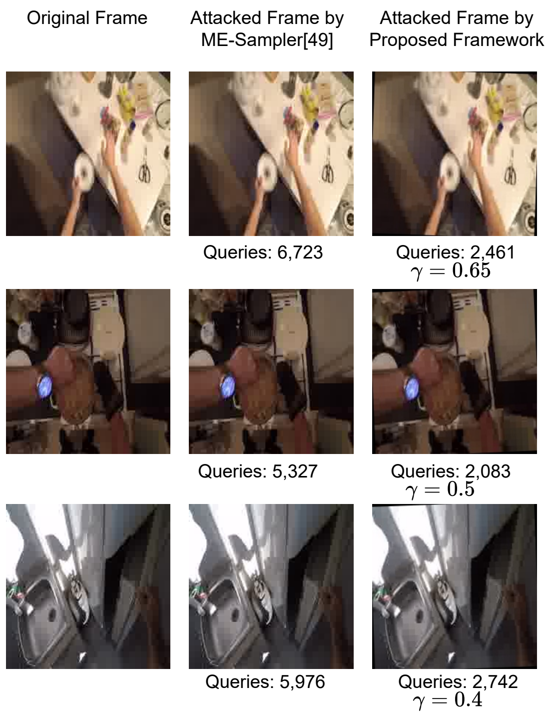

Figure 3. Comparative Qualitative Analysis of the proposed system. The detailed analysis is in the supplementary material. The first column shows the original frame, the second column shows the attacked frame using ME-Sampler [49], and the third column shows the attacked frame using our proposed technique.

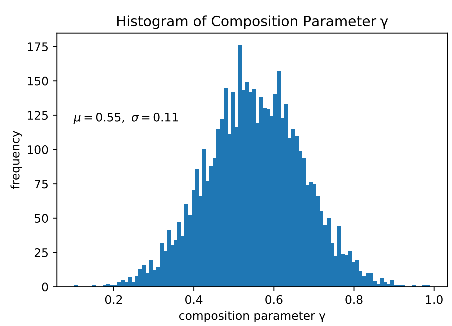

Figure 4. Histogram of the learned composition parameter on Epic-Kitchens dataset. The minimum and maximum values of γ are 0.07 and 0.96 respectively. Given such a variability of γ, learnable gradient composition is required for successful attacks.

is the learnable gradient composition framework, where the composition parameter γ is learned using DifferenceNet. Fig. 4 shows the histogram of the learned composition parameters on Epic-Kitchens dataset. We see that the distribution of γ parameter is similar to the Gaussian distribution. We report the mean of the Gaussian as 0.55 and standard deviation as 0.11. The minimum and maximum values of γ are 0.07 and 0.96 respectively. Given such a variability of γ (for successful attacks), it is no surprise that manual gradient composition fails completely as also shown in our ablation study (see Tab. 3).

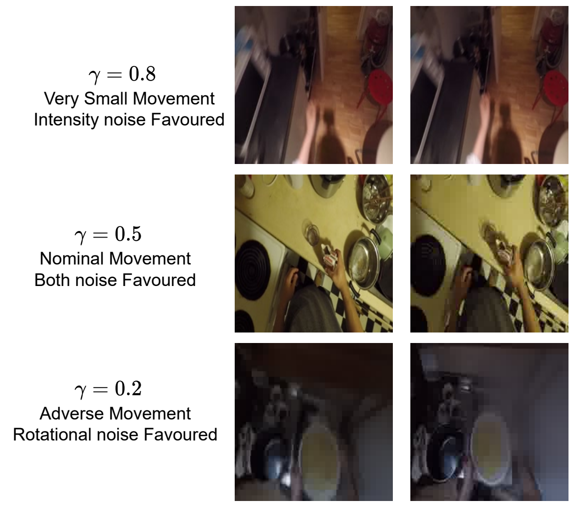

Figure 5. Videos having low, middle and high γ values. Videos having small spatio-temporal variation, high γ. Videos having large spatio-temporal variation, low γ.

Relationship between γ and Video Content: To understand the relationship between γ value and the corresponding video, we chose few videos having low, middle and high γ values. A few representation frames of these videos are shown in Fig. 5. We observe that the videos having small spatio-temporal variation, results in higher γ. Conversely, large variations results in smaller γ. This is expected, since in the videos where spatio-temporal variation is small, intensity-based noise has more affect rather than geometric noise. Hence, the proposed framework favors intensity noise by learning a high γ value.

5. Conclusion

Black-Box adversarial attacks on DNNs for videos analysis have utilized intensity-based noise for adversarial perturbation. However, such frameworks, require a large number of queries for estimating the perturbation. To overcome that, we propose a parametric noise based adversarial attack. It utilizes both intensity-based noise and camera rotational noise for generating the adversarial video. Gradient estimation has been done over both noises and are merged using a learnable novel gradient composition framework. We have shown the efficacy of the proposed framework on both first-person and third-person video analysis tasks.

6. Acknowledgement

This work was supported in part by the DST, Government of India, under project id T-138.

References

[1] Bharat Lal Bhatnagar, Suriya Singh, Chetan Arora, and CV Jawahar. Unsupervised learning of deep feature representation for clustering egocentric actions. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1447–1453, 2017. 2

[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. 2, 5

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[4] 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

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 5

[6] Jiawei Du, Hu Zhang, Joey Tianyi Zhou, Yi Yang, and Jiashi Feng. Query-efficient meta attack to deep neural networks. *arXiv preprint arXiv:1906.02398*, 2019. 1, 2

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2

[8] Jessica Finocchiaro, Aisha Urooj Khan, and Ali Borji. Egocentric height estimation. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1142–1150. IEEE, 2017. 2

[9] Antonino Furnari and Giovanni Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 6

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[12] 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. 1

[13] Joel A Hesch and Stergios I Roumeliotis. Consistency analysis and improvement for single-camera localization. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 15–22. IEEE, 2012. 2

[14] Yedid Hoshen and Shmuel Peleg. An egocentric look at video photographer identity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4284–4292, 2016. 2

[15] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020. 2

[16] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pages 2137–2146. PMLR, 2018. 1, 2

[17] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*, 2018. 1, 2, 4

[18] 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

[19] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 2

[20] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. Black-box adversarial attacks on video recognition models. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 864–872, 2019. 2, 6, 7

[21] Kris M Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR 2011*, pages 3241–3248. IEEE, 2011. 2

[22] 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

[23] Ana Cristina Murillo, Daniel Gutiérrez-Gómez, Alejandro Rituerto, Luis Puig, and Josechu J Guerrero. Wearable omnidirectional vision system for personal localization and guidance. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 8–14. IEEE, 2012. 2

[24] Pravin Nagar, Mansi Khemka, and Chetan Arora. Concept drift detection for multivariate data streams and temporal segmentation of daylong egocentric videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1065–1074, 2020. 2

[25] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017. 1, 2

[26] Hyun Park, Eakta Jain, and Yaser Sheikh. 3d social saliency from head-mounted cameras. *Advances in Neural Information Processing Systems*, 25:422–430, 2012. 2

[27] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. Predicting primary gaze behavior using social saliency fields. In *Pro-*

ceedings of the IEEE International Conference on Computer Vision, pages 3503–3510, 2013. 2

[28] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2847–2854. IEEE, 2012. 2

[29] Yair Poleg, Chetan Arora, and Shmuel Peleg. Temporal segmentation of egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2544, 2014. 2

[30] Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. Compact cnn for indexing egocentric videos. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. 2

[31] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 2

[32] Anuj Rathore, Pravin Nagar, Chetan Arora, and CV Jawahar. Generating 1 minute summaries of day long egocentric videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2305–2313, 2019. 2

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1

[34] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 2

[35] Suriya Singh, Chetan Arora, and CV Jawahar. First person action recognition using deep learned descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2620–2628, 2016. 2

[36] Suriya Singh, Chetan Arora, and CV Jawahar. Trajectory aligned features for first person action recognition. *Pattern Recognition*, 62:45–55, 2017. 2

[37] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2

[38] Daksh Thapar, Chetan Arora, and Aditya Nigam. Is sharing of egocentric video giving away your biometric signature? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 399–416. Springer, 2020. 2, 5, 6

[39] Daksh Thapar, Aditya Nigam, and Chetan Arora. Recognizing camera wearer from hand gestures in egocentric videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2095–2103, 2020. 2

[40] 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

[41] 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

[42] Sagar Verma, Pravin Nagar, Divam Gupta, and Chetan Arora. Making third person techniques recognize first-person actions in egocentric videos. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2301–2305. IEEE, 2018. 2

[43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

[44] 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

[45] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12338–12345, 2020. 2

[46] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2235–2244, 2015. 2

[47] Huanqian Yan, Xingxing Wei, and Bo Li. Sparse black-box video attack with reinforcement learning. *arXiv preprint arXiv:2001.03754*, 2020. 2

[48] Zhao Yang, Tie Liu, Jiehao Liu, Li Wang, and Sai Zhao. A novel soft margin loss function for deep discriminative embedding learning. *IEEE Access*, 8:202785–202794, 2020. 5

[49] Hu Zhang, Linchao Zhu, Yi Zhu, and Yi Yang. Motion-excited sampler: Video adversarial attack with sparked prior. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 240–256. Springer, 2020. 2, 4, 5, 6, 7, 8

[50] Linchao Zhu, Du Tran, Laura Sevilla-Lara, Yi Yang, Matt Feiszli, and Heng Wang. Faster recurrent networks for efficient video classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13098–13105, 2020. 2

[51] Linchao Zhu and Yi Yang. Label independent memory for semi-supervised few-shot video classification. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2020. 2