Efficient Decision-based Black-box Patch Attacks on Video Recognition

Kaixun Jiang[1*]   Zhaoyu Chen[1*]   Hao Huang[1]   Jiafeng Wang[2]
Dingkang Yang[1]   Bo Li[3]   Yan Wang[1†]   Wenqiang Zhang[1,2†]

[1]Academy for Engineering and Technology, Fudan University
[2]School of Computer Science, Fudan University   [3]vivo Mobile Communication Co., Ltd.

kxjiang22@m.fudan.edu.cn, {zhaoyuchen20, yanwang19, wqzhang}@fudan.edu.cn

Abstract

Although Deep Neural Networks (DNNs) have demonstrated excellent performance, they are vulnerable to adversarial patches that introduce perceptible and localized perturbations to the input. Generating adversarial patches on images has received much attention, while adversarial patches on videos have not been well investigated. Further, decision-based attacks, where attackers only access the predicted hard labels by querying threat models, have not been well explored on video models either, even if they are practical in real-world video recognition scenes. The absence of such studies leads to a huge gap in the robustness assessment for video models. To bridge this gap, this work first explores decision-based patch attacks on video models. We analyze that the huge parameter space brought by videos and the minimal information returned by decision-based models both greatly increase the attack difficulty and query burden. To achieve a query-efficient attack, we propose a spatial-temporal differential evolution (STDE) framework. First, STDE introduces target videos as patch textures and only adds patches on keyframes that are adaptively selected by temporal difference. Second, STDE takes minimizing the patch area as the optimization objective and adopts spatial-temporal mutation and crossover to search for the global optimum without falling into the local optimum. Experiments show STDE has demonstrated state-of-the-art performance in terms of threat, efficiency and imperceptibility. Hence, STDE has the potential to be a powerful tool for evaluating the robustness of video recognition models.

1. Introduction

Deep Neural Networks (DNNs) have showcased excellent efficacy in computer vision tasks [12, 47, 46, 25, 24, 6]. However, recent studies indicate they are vulnerable to ad-

*indicates equal contributions.
†indicates corresponding authors.

Figure 1. Given a source video and target video, STDE iteratively optimizes adversarial patches while requiring few queries. Compared to BSCA [4], STDE can achieve more potent attacks (e.g. targeted attacks) and generate smaller and sparser patches to improve the imperceptibility.

versarial examples, which are carefully crafted inputs with imperceptible adversarial perturbations. When facing adversarial examples [34], DNNs predict wrong outputs with high confidence, posing serious security threats. Most studies of adversarial examples focus on the image models [20, 14]. Wei et al. [39] indicate that the vulnerability also exists on video models. This vulnerability brings security threats on video-related tasks, such as object tracking [22, 23], video action recognition [19], etc. Therefore, it is indispensable to conduct thorough research on video adversarial examples for constructing more secure models.

Early studies of adversarial examples are concerned with perturbation-based attacks on image models [34, 20, 26, 2], where attackers add small l_p norm perturbations to each pixel of inputs. Recently, researchers have found that patch attacks, where attackers add regional and perceptible patches on inputs, pose a significant threat to DNNs. Further, many defense methods that are effective against perturbation-based attacks turn ineffective against patch at-

tacks [44]. This demonstrates that patch attacks are non-negligible for evaluating model robustness. Although patch attacks have made great progress on image models [1, 18, 11, 44, 9, 16, 7], to our best investigation, BSCA [4] is currently the only patch attack against video recognition models. Since BSCA focuses too much on the imperceptibility of the patch, its attack performance is relatively moderate (especially for targeted attacks). Therefore, it is of great interest to study a video patch attack that balances threat and imperceptibility to evaluate the robustness of video models.

In pioneering work [1, 18], patch attacks are carried out in the white-box setting, where attackers can access the whole details of the model (e.g. parameters and gradients). Since the white-box is too ideal, more work [11] has explored black-box score-based attacks, where the adversary can only access the model output (e.g. labels and corresponding logits) by querying models and utilize logits to design optimization strategies. Existing works propose rich patch forms based on applications, including monochrome [11], color texture [44, 9], watermark [16] and bullet screens [4], etc. However, as model security and privacy concerns continue to grow, obtaining logits has become increasingly difficult, particularly for video models used in security-critical tasks. Therefore, it is of practical significance for video models to investigate decision-based patch attacks, which can only access the predicted label of the model and pose a greater threat. However, to the best of our knowledge, there is currently a lack of research in this area. This gap leads to potential vulnerabilities for video models and impedes the advancement of developing more robust video models.

To bridge this gap, we explore a new attack setting called decision-based patch attacks on video models. This setting combines the advantage of patch and decision-based attacks to improve the assessment system for video model robustness. However, there also exists lots of challenges: 1) Compared to images, the temporal dimension of videos substantially enlarges the parameter space and incurs a significant query burden. Particularly, the mutual complement of information between frames increases the difficulty of attack. 2) The large parameter space of the patch (position, shape, texture) and the scarce output of the model (top-1 predicted label) can easily lead the attack to local optima [30], which reduces the efficiency of the attack.

To solve these challenges and achieve query-efficient decision-based patch attacks on video models, we propose **S**patial-**T**emporal **D**ifferential **E**volution patch attack (STDE). To improve query efficiency, STDE reduces the parameter space in the temporal and spatial domains. Specifically, in the spatial domain, STDE introduces target videos as prior knowledge to fill the texture of the patch and uses paired coordinates to model the position and shape of the patch. In the temporal domain, STDE performs binary en-coding on the video sequence and selects keyframes according to the temporal difference, achieving a sparse attack. After the modeling described above, the decision-based patch attack is transformed into a discrete combinatorial optimization problem. Therefore, we improve the classic heuristic algorithm [32] and propose a spatial-temporal differential evolution strategy that uses spatial-temporal mutation and crossover operations to avoid local optima. Figure 1 shows an illustration of generating the video patch based on STDE in the decision-based patch attack setting. Our main contributions and experiments are as follows:

- To our best knowledge, We are the first to combine patch attacks and decision-based attacks on video models, introducing a new attack scenario called decision-based patch attacks. This bridges the gap in robustness assessment of video models.

- To achieve query-efficient attacks in this new setting, we propose a novel attack STDE, which reduces the parameter space of video adversarial examples in spatial and temporal domains and uses spatial-temporal difference to search for the decision boundary.

- We conduct extensive experiments on video recognition models trained with UCF-101 and Kinetics-400 datasets. Compared with state-of-the-art methods, STDE shows 100% fooling rates with smaller patch area and fewer queries. Due to the sparse distribution and small size of adversarial patches, STDE ensures both potent attack capability and imperceptibility.

2. Related Work

2.1. Black-box Patch-based Attacks on Images

Patch-based attacks mislead DNNs by adding perceptible patches to the local region of the input [1, 18, 8], which can break the defense methods against perturbation-based attacks and can be well applied in real-world scenes. The parameters of the patch include texture, position, and shape [7], and the complex parameter space makes it difficult to optimize. Hastings Patch Attack (HPA) [11] is the first black-box patch attack, using the monochrome patch to fill texture and optimizing position and shape by Metropolis-Hastings sampling. Then, to improve query efficiency, Monochrome Patch Attack (MPA) [44] uses the reinforcement learning framework to optimize the position and shape of the patch. Due to the poor attack performance of monochrome patch for targeted attacks, Texture-based Patch Attack (TPA) [44] generates a texture dictionary as the texture of the adversarial patch based on training data. This prior knowledge increases the attack performance for targeted attacks significantly. It is, however, difficult to obtain the training data of the black-box model in practice.

Adv-watermark (AdvW) [16] optimizes the patch by the basin hopping evolution algorithm and converts the patch into the form of a watermark. Patch-Rs [9] adds small patches to the large patch region by random search. All of these attacks focus on images and ignore the temporal dimension. Therefore, transferring these methods to video will exponentially increase the query cost and attack difficulty. Moreover, these attacks depend on model feedback scores (i.e. logits) to craft adversarial patches, which can be shown in Table 1 to have poor attack performance in the decision-based setting.

2.2. Black-box Attacks on Videos

V-BAD [17] is the first video black-box attack method, which initializes perturbations with a pre-trained model on the image and then rectifies perturbations with NES [15]. ME-Sampler [49] uses motion map as prior information to guide the initialization of perturbations, which has higher query efficiency than V-BAD. GEO-TRAP [21] reduces the parameter space by geometric transformation, and enables an efficient attack. The above methods add perturbations to the whole video, furthermore, these methods all estimate gradients by continuous scores to achieve effective attacks, but they do not work well in decision-based settings. In addition, these methods are perturbation-based methods and cannot be applied to patch attacks.

Other black-box video attack methods select keyframes to achieve sparse attacks, where attackers add perturbations to keyframes. Wei et al. [40] determine keyframes according to the score of query feedback by masking video frames one by one as input. However, it requires a lot of queries to determine whether it is a keyframe. Recent work [38, 37, 43] introduces reinforcement learning framework to select keyframes and updates agents parameters based on query feedback score information. Deep-SAVA [27] uses Bayesian Optimization to identify the most influential frames on the video, but still needs the feedback score to evaluate the quality of the selected keyframes. Although the existing attacks provide many ways to select keyframes, they all exploit the score of the model output to guide optimization. When the probability value is converted to a hard label, these attacks can greatly increase the query burden. Instead of relying on model feedback information for optimization, our proposed temporal difference aims at optimizing patch area in the temporal domain, utilizing temporal mutation and crossover to find the optimal distribution of keyframes.

BSCA [4] proposes a score-based patch attack on video recognition, treating the patch as bullet screens. However, its reinforcement learning framework and patch form exhibit significant drawbacks in terms of memory and time consumption (as shown in Figure 4), and its incapacity to carry out targeted attacks. In contrast, we first investigate the more challenging decision-based attack setting and generate more threatening and sparse video adversarial patches with low cost and high efficiency.

2.3. Evolutionary Algorithm in Adversarial Attack

Many black-box attack scenarios can be transformed into gradient-free optimization problems, making evolutionary algorithms (EA) effective in various attack scenarios. Previous works have successfully applied EA to score-based sparse attacks [33] and score-based patch attacks [16]. Furthermore, EA has been explored in decision-based scenarios with multiple attack forms [35, 10, 5]. Additionally, EA has been used in score-based video attacks [48], However, [33, 16, 48] consider maximizing loss as optimization objectives based on the score of model output, while [35, 10, 5] neither consider the characteristics of the video in modeling parameters nor the evolutionary process.

3. Methodology

3.1. Preliminary

We denote the clean video as $x \in \mathbb{R}^{T \times H \times W \times C}$ and its corresponding label $y \in Y = \{1, 2, \cdots, K\}$, where T, H, W, C, K denote the length of frames, height, width, channel, the number of classes respectively. We denote the video recognition model as $F(\cdot)$. In the decision-based setting, we can not get the internal information of $F(\cdot)$ and $F(\cdot)$ only outputs the top-1 class $\tilde{y} \in Y$, i.e. $F(x) = \tilde{y}$. In the patch attack setting, we introduce the video adversarial patch added on the clean video frames to fool video recognition models. It is composed of the perturbation $\delta \in \mathbb{R}^{T \times H \times W \times C}$ and the mask matrix $M \in \{0,1\}^{T \times H \times W}$. δ determines the texture of the patch and M determines the numbers, shape, and position of video patches. Adversarial video x_{adv} is represented:

$$x_{adv} = \sim M \odot x + M \odot \delta, \tag{1}$$

where $\sim$ means the bitwise inversion on M. $M^{(t)}(i, j) = 1$ denotes the pixel with the coordinates (i, j) on the tth frame is in the region of the adversarial patch.

Combining patch-based attack and decision-based settings, our attack goal considers two optimization problems. On the one hand, we hope x_{adv} could fool the video recognition model $F(\cdot)$ successfully, i.e. $F(x_{adv}) \neq y$ for untargeted attacks or $F(x_{adv}) = y_{adv}$ for targeted attacks, where y_{adv} represents the target label we assign and is not equal to y. On the other hand, we optimize the adversarial patch area to be as small as possible. The patch area is constrained by the l_0 norm $|| \cdot ||_0$. Our attack (e.g. targeted attack) can be formulated as follows:

$$\arg \min_{M, \delta} \|x - x_{adv}\|_0, \quad s.t. \quad F(x_{adv}) = y_{adv}. \tag{2}$$
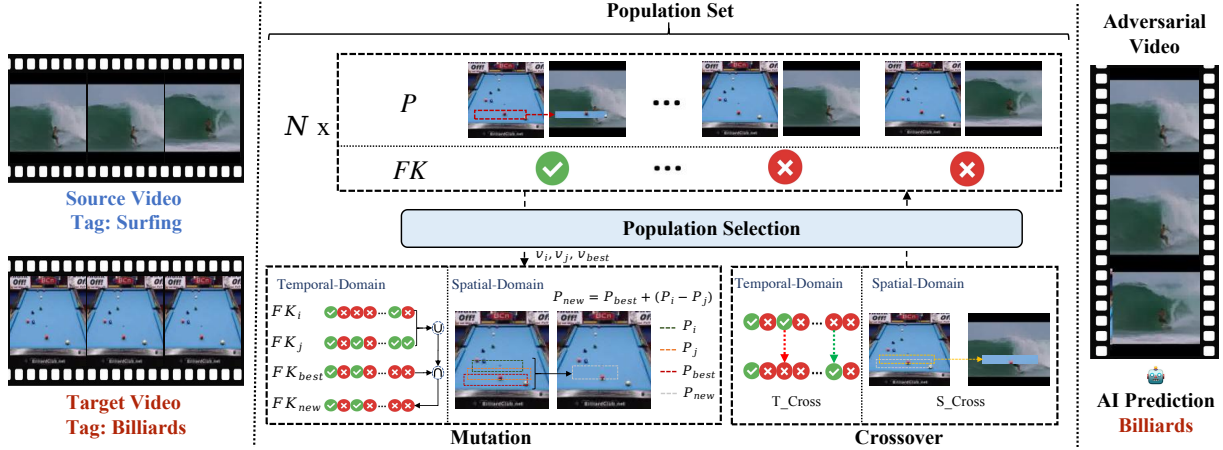
Figure 2. Overview of our decision-based video patch attack STDE. Given a clean video labeled *Surfing* and a target video labeled *Billiards*. STDE generates N populations through population initialization. Every population v is composed of the position set of patches P and keyframe binary sequence FK. Next, we randomly select v_i, v_j and the best population v_{best} to generate new population v_{new} by mutation and crossover in the spatial and temporal domains. Then, we use the fitness function to judge whether to introduce v_{new}. The final adversarial video can be predicted *Billiards* with minimal patch area.

3.2. Sparse Parameter Space

Spatial Domain. Generating a video adversarial patch needs two parameters: M and δ, where the former defines position and shape and the latter defines texture. For M, following the idea of generating the bounding box, we parameterize the coordinates of the upper left and lower right corners of the patch to generate a rectangular patch. For δ, the original parameters are RGB values of all pixels in the patch area, which are huge and difficult to optimize. Inspired by texture dictionaries [44], we introduce the target video x_{tar} as prior knowledge to fill patch texture instead of δ. In this way, we do not need any extra parameters. Compared with texture dictionaries generated by training data, x_{tar} has more explicit semantic information for targeted attacks and can be more easily accessed. We investigate the contribution of target videos to STDE and the robustness of STDE to different target videos and different source data in Section 4.3 and ablation studies. Therefore, x_{adv} generated by x_{tar} can be represented as follows:

$$x_{adv} =\sim M \odot x + M \odot x_{tar}. \tag{3}$$

For the targeted attack, target class becomes the label of the target video y_{tar}.

Temporal Domain. To further reduce the parameters, we select keyframes in the temporal domain and only add adversarial patches on keyframes. Specifically, we perform binary encoding on the video sequence and select keyframes by temporal difference. Temporal difference in our framework consists of temporal mutation and temporal crossover, which will be detailed described in the next section. Furthermore, temporal difference generates sparsely

distributed adversarial patches, enhancing imperceptibility.

3.3. Spatial-temporal Differential Evolution

The pipeline of STDE can be summarized as follows: first, we maintain a population set V containing N populations. Then, we select the optimal local population as the starting search direction and finally approach the optimal global population through spatial-temporal difference. The overall algorithm is shown in Algorithm 1, and the overall framework is illustrated in Figure 2.

Population Set. Every population v in V is composed of two parts: (P, FK), where P denotes the position of patches and FK denotes keyframe binary boolean vector. For P, $P = \{(p^{(0)}, p^{(1)}), (p^{(2)}, p^{(3)})\}^T$. Each frame needs two point pairs $\{(p^{(0)}, p^{(1)}), (p^{(2)}, p^{(3)})\}$, denoting coordinates of the upper left and lower right corner of each patch. These four parameters define the position and shape of the patch. For FK, $FK \in \{0,1\}^T$ and $FK^{(t)} = 1$ denotes the tth frame is a keyframe. Mask matrix M is generated by P and FK. Therefore, each population v corresponds to a mask matrix M. M and x_{tar} generate video adversarial examples x_{adv} according to Eq. 3.

Population Initialization. In our STDE framework, our attack starts from adversarial examples which can mislead threat models. Therefore, the population v that remains in the population set V must ensure that the generated x_{adv} is adversarial. For N initial populations, we sample initial P and FK from the uniform distribution. To initialize the population efficiently, P and FK are restricted by initialization rate μ and frame coverage rate cf which constrain the size and number of randomly generated patches, respectively. Pseudocode for population initialization is shown in

Supplementary Material. After population initialization, we get population set V, fitness set G which stores the fitness of each population v in V, and query cost q.

Fitness function. Score-based methods rely on the continuous score of query feedback to craft adversarial examples, which will lead to low optimization speed or even failure when the query feedback is replaced with discrete hard-label information. To achieve efficient attacks in the decision-based setting, we use the fitness function as new query feedback information. The fitness function evaluates the quality of each population in the evolutionary algorithm. Considering the optimization objective in Eq. 2, we introduce the patch area into the fitness function, and for the model query feedback information, we quantify it as 0 or ∞, corresponding to successful or unsuccessful attacks, respectively. After this quantification, we convert discrete model feedback information into continuous fitness function feedback, which accelerates the evolution speed. Fitness function can be expressed as (e.g. targeted attack):

$$g(x_{adv}) = \begin{cases} \|x - x_{adv}\|_0 - \lambda \cdot \mathrm{I}_t, & \text{if } F(x_{adv}) = y_{adv}, \\ \infty, & \text{otherwise,} \end{cases} \quad (4)$$

where $g(\cdot)$ denotes fitness function and λ denotes a balance weight. The first part of $g(\cdot)$ is to constrain the patch area by l_0 norm. The second is I_t, which calculates the intersection area between different patches of keyframes in the temporal dimension. The aggregated patches have more temporal semantic information than scattered patches, improving efficiency for targeted attacks. The effect of I_t is shown in Supplementary Material. Note that lower $g(\cdot)$ means better population.

Mutation. We generate new population by spatial and temporal mutation. First, we select v_i, v_j, v_{best} from population set, where v_i, v_j represent two populations randomly sampled from population set V. v_{best} denotes the population which has the smallest fitness value. $v_i \neq v_j \neq v_{best}$. Second, we take the optimal solution v_{best} of the current population set as the reference direction to search for the optimal global solution. It not only improves the query efficiency but also ensures the heritability of the new population. The process of mutation can be formulated as follows:

$$\begin{aligned} P_{new} &= P_{best} + \gamma \cdot (P_i - P_j), \\ FK_{new} &= FK_{best} \wedge (FK_i \vee FK_j). \end{aligned} \quad (5)$$

The patch represented by P belongs to the integer field, and integer operation is adopted to spatial mutation. γ denotes a constant as the mutation rate to control the differential variation [32]. FK denotes the keyframe binary boolean vector, using the binary operation to temporal mutation. Another advantage of temporal mutation is that the optimal solution tends to be searched in the direction with fewer keyframes, which improves the evolution speed.

Crossover. Crossover's role is to improve population diversity. As the iteration number increases, mutation tends

---

Algorithm 1 STDE for Targeted Attack
***
Input: video recognition model $F(\cdot)$, clean video x, ground truth y, target video x_{tar}, target video label y_{tar}
Parameter: mutation rate γ, crossover rate α, queries Q
Output: x_{adv}

1: $V, G, q = \text{Population\_init}(F, x, y, x_{tar}, y_{tar})$.
2: **while** $min(G) > \epsilon$ and $q < Q$ **do**
3:     $best, worst = \text{argmin}(G), \text{argmax}(G)$.
4:     Random select v_i, v_j from V and $v_i \neq v_j \neq v_{best}$.
5:     Create P_{new}, FK_{new} using Eq. 5.
6:     $P_{new} = \text{S\_Cross}(P_{new}, \gamma)$.
7:     $FK_{new} = \text{T\_Cross}(FK_{new}, \alpha)$.
8:     $v_{new} = (P_{new}, FK_{new})$.
9:     Generate M using v_{new}.
10:    Generate x_{adv} using Eq. 3 with x, x_{tar}, M.
11:    Calculate $g(x_{adv})$ using Eq. 4.
12:    **if** $g(x_{adv}) < G_{worst}$ **then**
13:       $G_{worst} = g(x_{adv}), v_{worst} = v_{new}$.
14:    **end if**
15:    $q = q + 1$.
16: **end while**
17: $best = \text{argmin}(G)$.
18: Generate M using v_{best}.
19: Generate x_{adv} using Eq. 3 with x, x_{tar}, M.
20: **return** x_{adv}

---

to fall into the optimal local region near the previous optimal solution. Crossover can make it jump out of the local optimum to find better solutions. S_Cross denotes sparsity crossover. For each element in P_{new}, we randomly add a noise k where $k \in \{-1, 0, 1\}$. To ensure the same degree of variation as mutation, we use the same γ to adjust. We describe temporal crossover as T_Cross. We randomly select α positions of FK_{new} to change the status of frames. α denotes a constant as crossover rate to control the degree of crossover. By controlling the magnitude of crossover through this constraint limit, the overall consistency of the optimization direction can be ensured.

Population Selection. If $g(v_{new}) < g(v_{worst})$, v_{new} will join population set V, the worst population v_{worst} will be eliminated. Otherwise, the population set is maintained.

4. Experiments

In this section, we conduct experiments compared with state-of-the-art methods. Then, we conduct diagnostic experiments, including hyper-parameters, target video source and image quality assessment. Moreover, we provide more ablation studies about STDE (e.g. crossover, fitness, and the robustness to target videos). Finally, we analyze time and memory costs, attack performance on patch defenses, and effectiveness analysis.

4.1. Experimental Settings

Datasets. We choose two popular datasets for video recognition: UCF-101 [31] and Kinetics-400 [3]. UCF-101 contains 13,320 video clips in 101 categories. Kinetics-400 includes 400 categories, with about 240,000 video clips for training and about 20,000 video clips for validation.

Video Recognition Models. We choose three popular models as our threat models, which show good performance on UCF-101 and Kinetics-400: C3D [12], Non-local (NL) [36] and TPN [45]. The input of all models except NL on Kinetics-400 is 16 consecutive frames randomly sampled from the video. To show the insensitivity of our method to the number of input frames, the input of NL on Kinetics-400 is 32 consecutive frames. On UCF-101, the accuracies for C3D, NL, and TPN are 86.3%, 74.4%, and 84.1%, respectively, while on Kinetics-400, the accuracies are 54.3%, 74.8% and 77.3% respectively.

Video Sampling. Following [17, 42, 41], we randomly sample one video from each class of test set on UCF-101 and validation set on Kinetics-400 as clean videos, and follow the same procedure to select target videos. We guarantee the video recognition model can accurately classify each video and the target video corresponding to each clean video belongs to a different class.

Metrics. 1) Fooling rate (FR): the ratio (%) of videos that are successfully attacked. 2) Average occluded area (AOA): the percentage (%) occluded by patches in total video area. 3) Average occluded area in the salient region (AOA*): the percentage (%) occluded by patches in the salient region, followed by [4]. 4) Average query number (AQN): average number of queries over all videos.

Baselines. Since we propose a decision-based patch attack on video models for the first time, there is no other method in line with this setting. Therefore, we compare four advanced patch attacks: TPA [44], AdvW [16], Patch-Rs [9], and BSCA [4]. These attacks are based on a score-based setting where attackers acquire more information from threat models than ours. To make a fair comparison, we convert score information into hard labels to meet the decision-based setting. To demonstrate the state-of-the-art performance of STDE, we retain the score-based BSCA (e.g. BSCA*) for comparison, although this comparison is unfair to our method. Following [44], we set the maximum allowed query number to 10,000 for untargeted attacks and 50,000 for targeted attacks. Since there are four metrics to compare performance, we guarantee almost the same AOA for each method to compare the remaining three metrics. More parameters are in Supplementary Material.

4.2. Performance Comparison

STDE vs Reinforcement Learning. To verify the superiority of our evolutionary algorithm in the decision-based setting, we compare STDE with TPA and BSCA, which utilize the reinforcement learning framework. TPA uses training data to generate texture dictionary to fill the patch texture. Because training data is not available in the black-box setting, we use target video instead of texture dictionary. It can be seen from Table 1 that our method is significantly better than TPA in all metrics.

BSCA is the first patch attack on video recognition. However, BSCA can not achieve the targeted attack. BSCA* in Table 1 denotes score-based BSCA. Experiments show that STDE performs better in all metrics than both BSCA and BSCA*. To demonstrate the superiority of STDE from another perspective and the portability of STDE, we design the white bullet screen attack (STDE-W) and the prior bullet screen attack (STDE-P) based on STDE. Different from STDE, the population parameters of STDE-W and STDE-P consist of the number of bullet screens, the coordinate of the upper left corner of bullet screens, and the transparency of bullet screens. The texture of STDE-W is filled with white, and the texture of STDE-P is filled with target videos. Experimental results are shown in Table 2. For the untargeted attack, FR of STDE-W is 6.9% higher than BSCA, and STDE-W only needs one-fourth of the query budget of BSCA. Compared with STDE-W, STDE-P has a higher fooling rate for the targeted attack. This also demonstrates target videos as prior knowledge can improve attack performance for targeted attacks.

STDE vs Random Search. An evolutionary algorithm is essentially an optimization search algorithm. To compare its advantages with random search, we compare it with Patch-Rs [9]. Patch-Rs uses random search to add small patches into large patch area and has strong attack performance in the image domain. From Table 1, we can see that our method can exceed Patch-Rs in terms of attack intensity and efficiency, on both untargeted and targeted attacks.

STDE vs Bash Hopping Evolution. To verify the superiority of STDE compared with other evolutionary algorithms, we compare with AdvW [16]. AdvW adopts the basin hopping evolution algorithm, a heuristic random search algorithm based on basin jumping. Experiments show that AdvW has the worst effect for the untargeted attack. The is because AdvW selects populations based on the probabilities fed by the model, which is difficult to optimize in the decision-based settings.

Overall. STDE has reached 100% FR on each model and dataset with 3% AOA on average for untargeted attacks and 15% AOA on average for targeted attacks. Among all tasks, NL/Kinetics-400 on targeted attacks has the lowest FR for others, our method can still improve 49.38 % FR with 1/7 queries. Overall, STDE achieves state-of-the-art performance in terms of threat, efficiency and imperceptibility.

Model	Method	UCF-101								Kinetics-400							
		Untargeted Attack				Targeted Attack				Untargeted Attack				Targeted Attack			
		FR↑	AOA↓	AOA*↓	AQN↓	FR↑	AOA↓	AOA*↓	AQN↓	FR↑	AOA↓	AOA*↓	AQN↓	FR↑	AOA↓	AOA*↓	AQN↓
C3D	TPA	56.40	4.98	2.77	5853	82.1	25.00	11.96	16485	73.50	3.85	1.94	3957	82.60	36.80	15.80	12617
	Patch-Rs	26.70	5.90	2.40	6567	3.96	25.00	10.12	47811	62.20	4.00	1.58	3970	2.26	37.00	15.05	48899
	AdvW	9.90	4.59	**1.40**	9090		Not Applicable			44.30	2.87	1.08	5944		Not Applicable		
	BSCA*	70.90	6.93	2.56	4222		Not Applicable			90.32	4.49	1.32	2217		Not Applicable		
	BSCA	59.40	6.31	2.31	4821		Not Applicable			81.10	4.20	1.15	3003		Not Applicable		
	Ours*	**100**	4.19	1.67	2830	**100**	24.90	10.40	2920	**100**	2.30	0.86	**1780**	**100**	36.30	13.70	8516
	Ours	**100**	**4.06**	1.64	**2600**	**100**	**18.70**	**7.20**	**2700**	**100**	**2.06**	**0.77**	1800	**100**	**26.10**	**9.70**	**8310**
NL	TPA	71.80	1.90	0.99	4839	58.41	13.72	5.85	24049	64.90	3.85	1.93	5033	51.62	18.75	7.70	27729
	Patch-Rs	48.50	1.90	**0.07**	5357	6.93	17.00	6.12	47373	48.30	3.99	1.31	5388	1.00	18.00	5.48	49604
	AdvW	39.60	1.91	0.48	6444		Not Applicable			36.59	2.87	0.83	6635		Not Applicable		
	BSCA*	78.20	3.67	0.93	3405		Not Applicable			73.68	3.82	0.95	4252		Not Applicable		
	BSCA	68.30	3.37	0.86	5059		Not Applicable			60.65	3.48	0.82	6820		Not Applicable		
	Ours*	**100**	1.12	0.40	2670	**100**	12.00	4.22	2920	**100**	2.71	0.80	2941	**100**	18.40	5.45	3896
	Ours	**100**	**0.70**	0.26	**2620**	**100**	**9.33**	**3.20**	**2820**	**100**	**1.03**	**0.34**	**2783**	**100**	**7.02**	**2.19**	**3751**
TPN	TPA	66.30	3.85	1.98	5196	80.19	13.70	5.99	14514	70.90	6.93	3.39	4234	71.67	20.10	8.33	18330
	Patch-Rs	64.35	5.00	1.78	3935	4.95	13.00	4.39	47937	54.80	7.00	2.30	4815	1.20	23.00	7.37	49473
	AdvW	40.60	3.85	**1.01**	6400		Not Applicable			49.12	6.93	2.06	5356		Not Applicable		
	BSCA*	82.10	4.26	1.20	3151		Not Applicable			64.91	7.19	2.09	5966		Not Applicable		
	BSCA	65.30	3.91	1.02	5724		Not Applicable			67.60	7.21	2.05	5597		Not Applicable		
	Ours*	**100**	**3.47**	1.33	2909	**100**	10.90	3.46	2985	**100**	6.30	**1.95**	**2284**	**100**	22.10	6.66	3872
	Ours	**100**	3.64	1.41	**2730**	**100**	**9.41**	**3.08**	**2850**	**100**	**6.14**	2.01	2798	**100**	**20.00**	**6.34**	**3677**

Table 1. Comparison of attack performance of STDE to other methods over UCF-101 and Kinetics-400 datasets using C3D, NL, and TPN.

Method	Untargeted				Targeted			
	FR↑	AOA↓	AOA*↓	AQN↓	FR↑	AOA↓	AOA*↓	AQN↓
BSCA	59.4	6.31	2.31	4821		Not Applicable		
STDE-W	**65.3**	**6.28**	2.45	**1229**	17.8	23.40	**9.25**	3410
STDE-P	60.4	6.30	**2.29**	1773	23.8	23.40	9.69	**3265**

Table 2. Two bullet screen attacks based on STDE (STDE-W and STDE-P) compared with BSCA on UCF-101 dataset against C3D model in the decision-based setting.
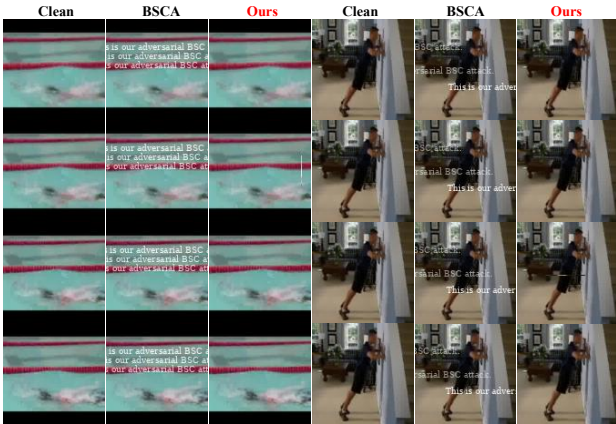
Model	Untargeted				Targeted			
	FR↑	AOA↓	AOA*↓	AQN↓	FR↑	AOA↓	AOA*↓	AQN↓
C3D	100	4.28	1.80	2690	100	19.00	7.69	2750
NL	100	0.55	0.20	2516	100	5.32	1.76	2800
TPN	100	2.97	1.07	2674	100	6.80	2.40	2886

Table 3. Attack performance with different source data.

Figure 3. We visualize our method compared with BSCA on UCF-101 dataset for the untargeted attack against C3D model.

4.3. Diagnostic Experiments

Hyper-parameters Tuning. We randomly select one video from each category on the UCF-101 test set, which does not overlap with the videos in Table 1. We select C3D as the threat model. The hyper-parameters of STDE are: $N = 15, \mu = 0.4, \gamma = 1, \lambda = 1.0$. For the untargeted attack, $cf = 0.6, \alpha = 1$. For the targeted attack, $cf = 0.7, \alpha = 2$. Please refer more details to Supplementary Material.

Target Video Source. To verify the generalization of STDE, we set 10 videos from the YouTube website as target videos. Clean videos are selected from UCF-101. Experiments in Table 3 show that different source data has little effect, indicating that STDE is robust to target video source.

Quality Assessment. Although patch attack is a form of perceptible attack that does not emphasize imperceptibility [1, 18], STDE still maintains its superiority. Figure 3 shows visualizations of STDE and BSCA. STDE has better imperceptibility due to small AOA and sparse distribution of patches. Moreover, following perturbation-based attacks and BSCA, we adopt PSNR, SSIM, MSE and AOA* to quantify the quality of generated adversarial videos. Table 4 shows STDE outperforms BSCA in all metrics. For further visualizations and analysis, please refer to Supplementary Material.

Method	SSIM↑	PSNR↑	AOA*↓	MSE↓
BSCA	0.864	22.74	2.310	430.3
STDE	**0.978**	**34.04**	**1.640**	**120.2**

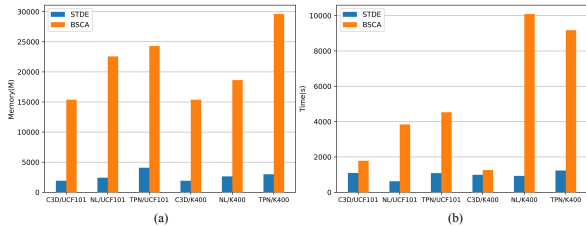Table 4. Quality assessment of adversarial videos generated by STDE and BSCA against C3D model.

Figure 4. Comparison with BSCA about memory (a) and time (b) costs under a single video attack with 10,000 queries.

4.4. Ablation Studies

Temporal Difference. We verify the role of keyframe selection with temporal difference. Ours* in Table 1 indicates STDE only performs spatial differential evolution. The experimental results show the temporal difference module can improve our method's attack performance, especially for targeted attacks.

Fitness Function. We explore the effects of different norm constraints on the performance, and Figure 5(a) shows the effects of l_0 and l_2 norms on untargeted and targeted attacks. The advantage of l_0 norm is more significant for targeted attacks, with AOA dropping by 1%, compared to untargeted attacks where AOA drops by 0.16%.

Crossover. Figure 5(b) shows the impact of crossover on attack performance (AOA). Without crossover, STDE falls into a local optimum, which leads to premature convergence, especially for targeted attacks (e.g. AOA increases by 20%).

Different Types. To verify the advantage of target video, we introduce two other forms of targets (Gaussian noise and Monochrome blocks) for comparison. Figure 5(c) shows the performance comparison under untargeted attacks. Both Gaussian noise and Monochrome blocks result in a loss of performance (e.g. 5% increase in AOA). In addition, for targeted attacks, the attack performance is further limited by the inability of Monochrome blocks and Gaussian noise to successfully initialize the population.

Different Videos. To verify the robustness of STDE to different target videos, we design five sets of experiments on C3D/UCF-101, each with the same 50 source videos and 1 different target video. Figure 5(d) shows the effect of different target videos on the untargeted attack with almost no effect and a slight effect on the target attack.

Model	Method	Untargeted				Targeted			
		FR↑	AOA↓	AOA*↓	AQN↓	FR↑	AOA↓	AOA*↓	AQN↓
C3D	Clean	**100**	4.06	1.64	2600	**100**	**18.70**	**7.20**	2700
	LGS	**100**	**2.96**	**1.25**	2730	**100**	19.00	7.87	2800
	DW	**100**	4.00	1.61	**2523**	**100**	18.96	7.94	**2557**
NL	Clean	**100**	0.70	0.26	2620	**100**	9.33	3.20	2820
	LGS	**100**	**0.62**	**0.23**	2370	99	**9.02**	**3.11**	2760
	DW	98	0.63	**0.23**	**2349**	98	12.45	4.27	**2672**
TPN	Clean	**100**	3.64	1.41	**2730**	**100**	9.41	3.08	2850
	LGS	**100**	**2.64**	**1.04**	2810	**100**	**8.59**	**2.90**	2850
	DW	**100**	3.33	1.28	2760	**100**	10.15	3.58	**2848**

Table 5. Attack performance against LGS and DW. Clean denotes no defense methods to be used.

4.5. Time and Memory Cost

To evaluate the attack cost of video patch attacks, we analyze the time and memory costs for 10,000 single-video attacks between STDE and BSCA on a single V100 GPU. Figure 4 shows that under all models and datasets, our method has a higher query efficiency while requiring only one-seventh of the memory cost compared to BSCA. We analyze the space and time complexity of STDE and BSCA. The space complexity is $O(1)$ and $O(n)$, respectively. Time complexity is $O(T \times Q)$ and $O(m \times T \times Q)$, where m is the number of bullet screens.

4.6. Attacks on Patch Defenses

To verify the robustness of STDE, we select two state-of-the-art patch defense methods. Local Gradient Smoothing (LGS) [28] defends by inhibiting the values of high activation areas because they consider the added patch has a high activation value. Digital Watermarking (DW) [13] defends by eliminating the watermark. We apply the above methods to every frame for achieving video defense. Table 5 shows neither of these defense methods can defend against STDE. We observe that the clean video is damaged to some extent due to these two pre-processing defense methods, improving attack performance instead.

4.7. Effectiveness Analysis

We are the first to explore the feasibility of patch-based targeted attacks on video models. To explore how STDE achieves the targeted attack, we use Grad-Cam [29] to visualize the attention area in Figure 6. For clean videos, we observe that the model focuses on action scenes, which is consistent with human intuition. For adversarial videos, we discover that just covering a small area with adversarial patches on the clean video can cause attention to be transferred to the adversarial patch for targeted attacks, and these patches may not be added on consecutive frames. This suggests that generated adversarial patches fool the video recognition model by changing the attention distribution of the model and also exhibit high redundancy of video information in the spatial and temporal domains.
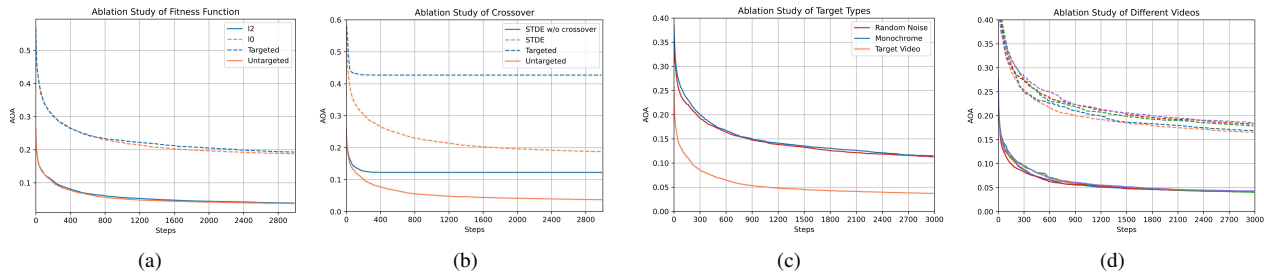
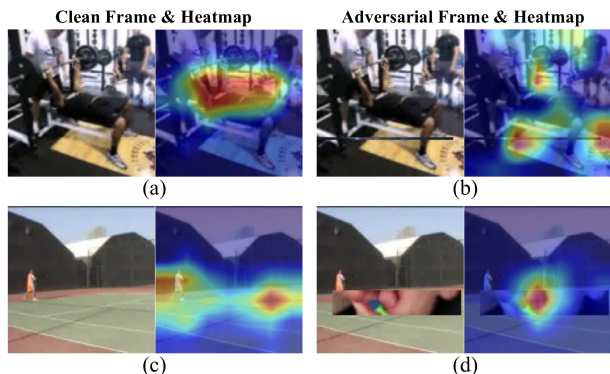Figure 5. Ablation studies on the UCF-101 dataset against C3D model.

Figure 6. (a) and (c) represent clean video frames & heatmaps. (b) and (d) represent untargeted and targeted adversarial video frames & heatmaps respectively.

5. Conclusion

In this paper, we study the vulnerability of video models in a new attack setting called decision-based patch attacks on video recognition models. To achieve query-efficient attacks in this new attack setting, we propose a spatial-temporal differential evolution algorithm (STDE) framework which is simple but effective. Extensive experiments show that STDE achieves state-of-the-art performance in terms of threat and imperceptibility with low cost and high query effiency. We further analyze the effectiveness and robustness of STDE. We also indicate that STDE can easily migrate to other forms of patch attacks. This work explores a new attack setting for video models and provides a powerful attack method in this setting which further improves the assessment system of video model robustness. In the future, we will make full use of the information from the video itself to explore more sparse and efficient video attacks in the decision-based setting.

Broader Impacts. Our work demonstrates that the vulnerability of DNNs exists not only for the image classification models but also for the video recognition models. This vulnerability causes security risks to video-related tasks in the real world. We propose STDE, as a powerful method to evaluate the robustness of video recognition models, to help researchers understand video recognition models from the perspective of adversarial attack, so as to design more robust video recognition models.

References

[1] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. In *Adv. Neural Inform. Process. Syst. Worksh.*, 2017. 2, 7

[2] 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. 1

[3] 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. 6

[4] Kai Chen, Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Attacking video recognition models with bullet-screen comments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 312–320, 2022. 1, 2, 3, 6

[5] Zhaoyu Chen, Bo Li, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. Query-efficient decision-based black-box patch attack. *arXiv preprint arXiv:2307.00477*, 2023. 3

[6] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. *arXiv preprint arXiv:2305.10665*, 2023. 1

[7] Zhaoyu Chen, Bo Li, Shuang Wu, Jianghe Xu, Shouhong Ding, and Wenqiang Zhang. Shape matters: Deformable patch attack. In *European conference on computer vision*, 2022. 2

[8] Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. Towards practical certifiable patch defense with vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15148–15158, 2022. 2

[9] Francesco Croce, Maksym Andriushchenko, Naman D Singh, Nicolas Flammarion, and Matthias Hein. Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6437–6445, 2022. 2, 3, 6

[10] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019. 3

[11] Alhussein Fawzi and Pascal Frossard. Measuring the effect of nuisance variables on classifiers. In *British Machine Vision Conference (BMVC)*, number CONF, 2016. 2

[12] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 1, 6

[13] Jamie Hayes. On visible adversarial perturbations & digital watermarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1597–1604, 2018. 8

[14] Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuze Zhang, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 989–997, 2022. 1

[15] 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. 3

[16] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Xiaoguang Han. Adv-watermark: A novel watermark perturbation for adversarial examples. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1579–1587, 2020. 2, 3, 6

[17] 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. 3, 6

[18] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning*, pages 2507–2515. PMLR, 2018. 2, 7

[19] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1725–1732. IEEE Computer Society, 2014. 1

[20] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 1

[21] Shasha Li, Abhishek Aich, Shitong Zhu, Salman Asif, Chengyu Song, Amit Roy-Chowdhury, and Srikanth Krishnamurthy. Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations. *Advances in Neural Information Processing Systems*, 34:2085–2096, 2021. 3

[22] Siyuan Liang, Xingxing Wei, Siyuan Yao, and Xiaochun Cao. Efficient adversarial attacks for visual object tracking. In *European Conference on Computer Vision*, pages 34–50. Springer, 2020. 1

[23] Siao Liu, Zhaoyu Chen, Wei Li, Jiwei Zhu, Jiafeng Wang, Wenqiang Zhang, and Zhongxue Gan. Efficient universal shuffle attack for visual object tracking. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2739–2743. IEEE, 2022. 1

[24] Yang Liu, Jing Liu, Kun Yang, Bobo Ju, Siao Liu, Yuzheng Wang, Dingkang Yang, Peng Sun, and Liang Song. Ampnet: Appearance-motion prototype network assisted automatic video anomaly detection system. *IEEE Transactions on Industrial Informatics*, pages 1–13, 2023. 1

[25] Yang Liu, Jing Liu, Mengyang Zhao, Shuang Li, and Liang Song. Collaborative normality learning framework for weakly supervised video anomaly detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69(5):2508–2512, 2022. 1

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1

[27] Ronghui Mu, Wenjie Ruan, Leandro Soriano Marcolino, and Qiang Ni. Sparse adversarial video attacks with spatial transformations. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 101. BMVA Press, 2021. 3

[28] Muzammal Naseer, Salman Khan, and Fatih Porikli. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1300–1307. IEEE, 2019. 8

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[30] Yucheng Shi, Yahong Han, Qinghua Hu, Yi Yang, and Qi Tian. Query-efficient black-box adversarial attack with customized iteration and sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[32] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997. 2, 5

[33] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE*

Transactions on Evolutionary Computation, 23(5):828–841, 2019. 3

[34] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014. 1

[35] Viet Quoc Vo, Ehsan Abbasnejad, and Damith Ranasinghe. Query efficient decision based sparse attacks against black-box deep learning models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[36] 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. 6

[37] Zeyuan Wang, Chaofeng Sha, and Su Yang. Reinforcement learning based sparse black-box adversarial attack on video recognition models. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3162–3168. ijcai.org, 2021. 3

[38] Xingxing Wei, Huanqian Yan, and Bo Li. Sparse black-box video attack with reinforcement learning. *Int. J. Comput. Vis.*, 130(6):1459–1473, 2022. 3

[39] 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. 1

[40] 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. 3

[41] Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Boosting the transferability of video adversarial examples via temporal translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2659–2667, 2022. 6

[42] Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Cross-modal transferable adversarial attacks from images to videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15064–15073, 2022. 6

[43] Huanqian Yan and Xingxing Wei. Efficient sparse attacks on videos using reinforcement learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2326–2334, 2021. 3

[44] Chenglin Yang, Adam Kortylewski, Cihang Xie, Yinzhi Cao, and Alan Yuille. Patchattack: A black-box texture-based attack with reinforcement learning. In *European Conference on Computer Vision*, pages 681–698. Springer, 2020. 2, 4, 6

[45] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 591–600, 2020. 6

[46] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, and Lihua Zhang. Context deconfounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19005–19015, June 2023. 1

[47] Dingkang Yang, Shuai Huang, Zhi Xu, Zhenpeng Li, Shunli Wang, Mingcheng Li, Yuzheng Wang, Yang Liu, Kun Yang, Zhaoyu Chen, et al. Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. *arXiv preprint arXiv:2307.13933*, 2023. 1

[48] Yu Zhan, Ying Fu, Liang Huang, Jianmin Guo, Heyuan Shi, Houbing Song, and Chao Hu. Cube-evo: A query-efficient black-box attack on video classification system. *IEEE Transactions on Reliability*, 2023. 3

[49] Hu Zhang, Linchao Zhu, Yi Zhu, and Yi Yang. Motion-excited sampler: Video adversarial attack with sparked prior. In *European Conference on Computer Vision*, pages 240–256. Springer, 2020. 3