StyleFool: Fooling Video Classification Systems via Style Transfer

Yuxin Cao*, Xi Xiao*, Ruoxi Sun†, Derui Wang†, Minhui Xue† and Sheng Wen‡
*Shenzhen International Graduate School, Tsinghua University, China
†CSIRO's Data61, Australia
‡Swinburne University of Technology, Australia

arXiv:2203.16000v3 [cs.CV] 2 May 2023

***Abstract*—Video classification systems are vulnerable to adversarial attacks, which can create severe security problems in video verification. Current black-box attacks need a large number of queries to succeed, resulting in high computational overhead in the process of attack. On the other hand, attacks with restricted perturbations are ineffective against defenses such as denoising or adversarial training. In this paper, we focus on unrestricted perturbations and propose *StyleFool*, a black-box video adversarial attack via style transfer to fool the video classification system. StyleFool first utilizes color theme proximity to select the best style image, which helps avoid unnatural details in the stylized videos. Meanwhile, the target class confidence is additionally considered in targeted attacks to influence the output distribution of the classifier by moving the stylized video closer to or even across the decision boundary. A gradient-free method is then employed to further optimize the adversarial perturbations. We carry out extensive experiments to evaluate StyleFool on two standard datasets, UCF-101 and HMDB-51. The experimental results demonstrate that StyleFool outperforms the state-of-the-art adversarial attacks in terms of both the number of queries and the robustness against existing defenses. Moreover, 50% of the stylized videos in untargeted attacks do not need any query since they can already fool the video classification model. Furthermore, we evaluate the indistinguishability through a user study to show that the adversarial samples of StyleFool look imperceptible to human eyes, despite unrestricted perturbations.**

***Index Terms*—video adversarial attack, video style transfer, unrestricted perturbations, black-box attack**

I. Introduction

With short videos becoming more and more popular in today's era, videos have gradually become inevitable in people's daily lives [1]. To date, a large number of video-sharing mobile applications have sprung up, such as Tiktok, Facebook Watch, and Youtube [2]–[4]. In order to avoid politically sensitive disputes and protect the physical and mental health of minors, it is necessary to verify the videos uploaded by users to prevent the spread of illegal or criminal videos such as violence, pornography, and malicious marketing [5], [6]. However, the rapid increase in the number of videos and the limitations of human and time resources have become challenges to manual verification. Therefore, the demand for machine-learning-assisted video classification systems increases greatly [5].

Severe consequences may occur once the video verification is compromised: an attacker can maliciously modify pixels of an illegal video, *e.g.,* a violent or pornographic video, to bypass the video verification classifier (*i.e.,* the video is falsely classified as benign) and expose the video to the public. If the illegal video is further circulated widely, it may cause public panic and other adverse effects, such as the threats to video content rating for children and the risk of video spreading for terrorist purposes in social networks [7], [8]. In addition, the emerging technologies in artificial intelligence, such as DeepFakes [9] and Face2Face [10], extremely reduce the difficulty to generate fake videos, making video verification more important than ever. This is substantiated by the recent fact that a fake video of the Ukrainian president calling on his soldiers to lay down their weapons was uploaded to a Ukrainian news website [11]. There is no doubt that the robustness of video classification systems is security-critical. It is essential for the security research community to investigate potential attacks against robustness and thoroughly evaluate the machine-learning-assisted system before full deployment.

As a family of machine learning, Deep Neural Networks (DNNs) have become an indispensable tool in the field of multimedia [22]–[26]. Despite significant advantages and far-reaching influence, studies have found that DNNs could be vulnerable to adversarial attacks [16], [27]. More specifically, by superimposing an elaborately designed perturbation (even unsuspicious or imperceptible to human eyes) on an input sample, an attacker can fool the classifier into misclassification [16], [28]. Recent research found that, as long as a pixel of the input image is perturbed, the classifier can be fooled successfully [29]. The resulting security problem has encountered enormous challenges with the application of DNNs in many aforementioned fields.

In contrast to a plethora of studies focusing on image adversarial attacks [7], [16], [28], [30]–[37], research in the video domain has been slowly ramping up [18]–[21], [38]–[41]. One of the main reasons is that videos contain temporal information, which greatly increases the attack difficulty. Table I shows the comparison of some existing video attacks, which can be categorized into two branches: *universal attacks* and *one-on-one attacks*. The attacks can be launched in either *online* scenario or *offline* scenario. The online scenario requires real-time attacks adding perturbations to online videos with minimal latency. In contrast, the offline scenario has a minimal concern towards the attacking overhead since the attacker has unlimited local access to the victim videos and models. Universal attacks such as C-DUP [18] and U3D [19] aim to generate a universal perturbation for all videos so as to fool the classifier. One-on-one attacks such as V-BAD [20] and H-Opt [21] aim to produce a sample-specific perturbation for each input video. Universal perturbations are only available in untargeted attacks, while one-on-one perturbations can be applied to both targeted and untargeted attacks. From the perspective of the adversary's knowledge, white-box attacks (*e.g.,* C-DUP [18]) have a stronger assumption that the adversary can access the

TABLE I: A comparison of adversarial attacks against video classifiers.

Approaches	Online	Offline	White-box	Black-box	Universal	One-on-one	Untargeted	Targeted	Restricted	Unrestricted	Compromised Defenses			
											AdvIT [12]	Comdefend [13], [14]	RS [15]	AT [16], [17]
C-DUP [18]	●	○	●	○	●	○	●	○	●	○	▒	▒	▒	▒
U3D [19]	●	○	○	●	●	○	●	○	●	○	■	▒	■	■
V-BAD [20]	○	●	○	●	○	●	●	●	●	○	□	□	□	▒
H-Opt [21]	○	●	○	●	○	●	●	●	●	○	□	□	□	▒
StyleFool (ours)	○ [1]	●	○	●	○ [1]	●	●	●	○	●	■	■	■	■

● : the item is supported by the attack; ○ : the item is not supported by the attack.
■ : the attack can compromise the defense; □ : the attack cannot compromise the defense; ▒ : not mentioned in the corresponding research.
[1] Although StyleFool proposed in this paper focuses on offline one-on-one attacks, we argue that our framework can also produce universal perturbations which is suitable for online attacks. See more details in the discussion.

training data and the classifier model. This makes white-box attacks not as practical as black-box attacks. Additionally, state-of-the-art video attacks, which introduce imperceptible ℓ_p-norm restricted perturbations, are shown to be successfully defended by adversarial defenses [42], such as AdvIT [12], Comdefend [13], [14], and Randomized Smoothing (RS) [15]. GAN-generated images can also be distinguished by CNN-generated image detection [43]. Furthermore, we point out that the existing one-on-one attacks [20], [21] consume a large number of queries (approximately $(5 \sim 26) \times 10^4$) when generating adversarial perturbations.

In summary, we list three limitations of the state-of-the-art video attacks as follows. (*i*) White-box attacks are less practical than black-box attacks for heavily overestimating the capability of the adversary. (*ii*) A large number of queries are needed in the existing video attacks [20], [21], resulting in high complexity especially when the target model is a commercial system. (*iii*) Restricted perturbations can be defended by adversarial defenses successfully.

In this research, we propose ***StyleFool***, a black-box unrestricted adversarial attack framework against video classification systems, that has the following advantages. (*i*) StyleFool introduces *unrestricted* perturbations which transfer videos into another style while preserving the semantic information in the black-box setting. Concretely, we focus on changing the non-semantic critical information which will not confuse human understanding of video content, but mislead the classifier. We argue such perturbations are observable but imperceptible that preserve the indistinguishability of original videos and break the limitation existing in the state-of-the-art attack frameworks. (*ii*) Style transfer can initialize powerful perturbations. Pre-processing videos by adversarial style transfer can significantly reduce the number of queries in the subsequent attack. (*iii*) Unrestricted perturbations with high temporal consistency are considered, which enhances the robustness against state-of-the-art defenses, including detection defense (*e.g.,* AdvIT [12]), reconstruction defense (*e.g.,* ComDefend [13]), certified defense (*e.g.,* RS [15]), and a CNN-generated detector [43]. The source code is publicly available at https://github.com/JosephCao0327/StyleFool.

Our main contributions are as follows.

- We propose a novel attack framework, StyleFool, in video domain. To the best of our knowledge, StyleFool is the *first* attempt to attack video classification systems with style-transfer-based unrestricted perturbations.
- We initialize perturbations with adversarial style transfer, which pushes the stylized video towards the decision boundary to reduce the queries required during adversarial sample generation. Meanwhile, the temporal consistency and indistinguishability (*i.e.,* whether an adversarial video can be distinguished from the original video by human subjects) of the video are preserved.
- Our extensive experiments indicate that StyleFool effectively moves the videos to the vicinity of decision boundaries. It also reduces the number of queries, compared to the state-of-the-art video attacks, V-BAD [20] and H-Opt [21], by 43% and 83%, respectively. A user study demonstrates the indistinguishability of StyleFool. We show that StyleFool is an efficient video attack framework, which is conducive to improving the robustness of the video classification model.
- We demonstrate the capability of StyleFool to bypass the state-of-the-art defense strategies, AdvIT [12], ComDefend [14], RS [15], and CNN-generated image detection [43]. We also discuss and recommend the potential mitigation. We believe that our research will arouse community's attention of such style-transfer-based attacks in security scenarios, especially in video verification.

Ethical considerations. The Human Research Ethics Committee of the authors' affiliation determined that the study was exempt from further human subjects review, and we followed the best practice for ethical human subjects survey research, *e.g.,* all questions were optional and we did not collect unnecessary personal information. All participants are over 18 years old and they consented for their answers to be used for academic research.

II. Background and Threat Model

In this section, we first introduce the background information and existing research related to our work. Then we describe the threat model from three aspects: adversary's goals, capabilities, and knowledge.

A. Background and Related Work

DNN-based video classification systems. A video classification system automatically classifies videos into corresponding labels based on the semantic content of human behaviors in complex events with DNNs, such as convolutional neural networks (CNNs). Several DNN-based video classification systems, such as TSN [44], C3D [45], LRCN [46], CNN+LSTM [47], and I3D [24], have been proposed for video recognition tasks. Among them, by incorporating temporal information into a two-dimensional CNN (2D-CNN) [44],

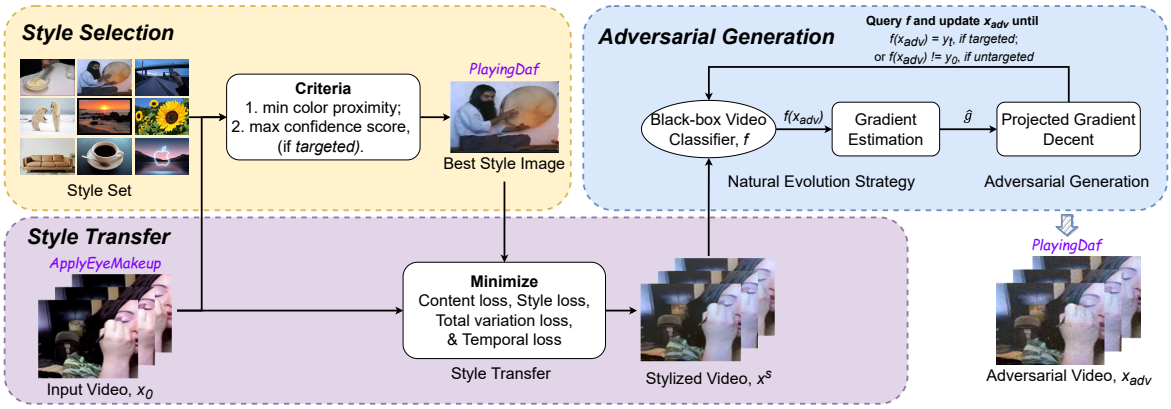


Fig. 1: An overview of StyleFool against video classification systems.

[48], a 3D-CNN network was proposed in C3D to extract spatiotemporal features of videos. Inspired by 3D ConvNets [23], TSN [44] and CNN+LSTM [47], I3D was proposed with two-stream inflated filters for video classification. C3D and I3D perform better than other models in video classification tasks and are prevalently adopted in various applications [24], [45]. On the flip side, the security issues in C3D and I3D are spotlighted due to their popularity [18]–[21].

Adversarial attacks and defenses. Existing studies in the field of adversarial attack mainly focus on the image domain. Most of them are based on ℓ_p-norm perturbations [7], [16], [28], [30]–[37], while some recent studies have considered unrestricted attacks, breaking the boundaries in image attacks [49]–[55]. For example, ColorFool [51] changed the color of the image and Patchattack [52] added a patch to the image. It has been verified that the attack performance of unrestricted attacks is better than that of restricted attacks [51], so as the circumvention performance against defenses [52].

Adversarial attacks against video classification can be classified as either white-box attacks [18], [38], [41], [56]–[58] or black-box attacks [19]–[21], [39], [40], [59]. The first attack towards video classifiers was operated in a white-box setting by optimizing perturbations bounded by ℓ_1 or ℓ_2 norm [38]. Inspired by Generative Adversarial Networks (GANs), an offline universal adversarial perturbation called C-DUP was proposed for white-box attacks [18]. One recent work was devoted to attacking the video compression model in the context of white-box setting, and the method was further exploited on the classification model for the adversarial attack [41]. V-BAD was proposed to launch attacks with limited queries against black-box classifiers [20]. Subsequently, in light of the Opt attack [36], a heuristic attack was proposed to increase the perturbation sparsity in black-box attacks by only adding perturbations to salient regions in key frames [21]. To enhance the transferability of black-box attacks, a universal three-dimensional perturbation (U3D) was lately proposed [19].

The state-of-the-art defenses detect or remove adversarial perturbations in the videos [12], [14]. AdvIT detects inconsistencies between video frames and frame-wise optical flows to expose adversarial videos [12]. Furthermore, Comdefend relies on compressing and reconstructing a video with latent noise to mitigate the effect of adversarial perturbations [14]. In the image domain, RS is proved to be effective for ℓ_2-norm attacks [15].

Video style transfer. Style transfer methods stylize a content image according to a style image, such as an artwork by a famous painter. Hertzmann *et al.* [60] first explored image style transfer using image analog. Subsequently, it was found that content loss and style loss can better preserve the similarity between an input image and the stylized image in the feature space [61]. Recently, the total variance loss was introduced to further ensure the smoothness of images at the pixel-wise level [62]. However, when applying style transfer to videos, the inconsistency between adjacent frames becomes a troublesome issue [63], [64]. To ease the problem, the temporal loss was further developed to generate consistent and stable stylized video sequences [65].

Adversarial attack using style transfer. AdvCam [53] first applied style transfer to image adversarial attacks by applying style transfer on specific areas in an image. However, in video scenarios, the most challenging point is to generate an adversarial stylized video while keeping the consistency of each frame in the video. What attacking video classifiers by style transfer differs from AdvCam is that, it is challenging to combine the loss of style transfer and adversarial attack, since the learning rates and the times of iteration in the two steps are rather different when learning video samples. Further, we cannot only select some specific areas in frames for video style transfer, considering that the style of an area is likely to change in a video (*e.g.,* the motion of an actor). We argue that it is difficult to squarely extend state-of-the-art style transfer attacks from images to videos.

B. Threat Model

We detail the threat model of StyleFool against video classification systems in this section.

Adversary's goals. Given a DNN $f(x) : x \rightarrow y$, it takes a video $x \in \mathbb{R}^{T\times H\times W\times C}$ from a video set X as input. Herein, T, H, W, and C are the frame number, height, width, and channel number of x, respectively. The model outputs a K-

class prediction label $y \in \{1, ...K\},\ y \in Y$, where Y is the set of predicted labels of X. The adversary's goal is to find an adversarial counterpart x_{adv} of x satisfying the constraints listed in Equation 1. The output label is a predetermined target class y_t in targeted attacks. Otherwise, in untargeted attacks, x_{adv} is misclassified into a random label other than the ground truth label y_0 of an input video x.

$$\begin{cases} f(x_{adv}) = y_t, & if\ targeted, \\ f(x_{adv}) \neq y_0, & if\ untargeted. \end{cases} \tag{1}$$

Adversary's capabilities. In our research, we limit the adversary's capabilities as follows. The adversary is capable of crafting perturbations and superimposing them on given videos in the offline setting. Please note that we only focus on one-on-one attacks which crafts sample-specific perturbations. Compared with universal attacks, one-on-one attacks can achieve both untargeted and targeted attacks.

To establish a video frame set from which the style images could be selected, the adversary may collect a large number of videos from publicly available datasets or other online resources. Furthermore, in the targeted attack scenario, when the style images from the targeted class are not satisfying, the adversary can search and download images or videos according to common sense about the target class to form the style set.

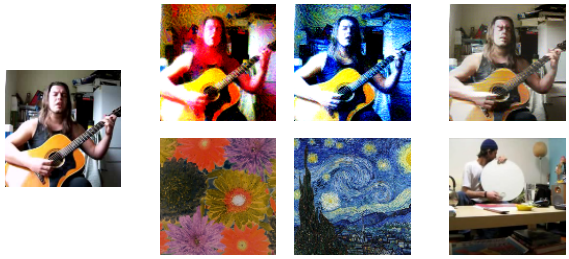
Adversary's knowledge. Different from the black-box setting in attacks towards images [34], [36], [66] and other state-of-the-art work in videos [21], we place stricter restrictions on the knowledge possessed by the adversary, *i.e.,* the adversary can neither access the attacked model parameters nor the training set. The adversary can only access the top-1 label y (the label with the highest confidence score) and its confidence score $p(y|x)$. An upper limit bound of attacker's budget on the number of queries is set; otherwise, in offline scenarios, the attack will always succeed if the query resources are not limited. Such a query limit also simulates the real-world black-box attacks, especially when the attacking targets are commercial video classifiers. Specifically, we adopt the *query-limited partial information* setting [35] in our research.

III. Design of StyleFool

The framework of StyleFool is depicted in Figure 1. StyleFool can be divided into three stages which are style selection, style transfer, and adversarial sample generation. At first, according to the style selection criteria (*i.e.,* the color theme proximity and the target class confidence), the best style is selected for style transfer. Then, by considering content, style, total variance, and temporal loss terms, the clean video is transferred into the selected style. Finally, a black-box adversarial attack is conducted to consolidate perturbations that can fool the target model.

A. Style Selection

Style selection serves for two purposes. First, it searches for style images that ensure the stylized videos are indistinguishable from the original videos. Moreover, properly selected style images help initialize the stylized video in the vicinity of the decision boundary to reduce the number of queries in the subsequent attacking process. As shown in Figure 2, the existing style transfer approaches [63]–[65] aim to conduct style transfer based on artistic styles that are distinct from the styles of the original images. Without a careful selection of style images, style transfer may produce changes that can be easily captured by human eyes. This is in contrast with the aim of adversarial attacks regarding the indistinguishability of perturbations. To preserve the indistinguishability of the stylized videos, we propose a collective set of style selection approaches for both targeted and untargeted attacks.



(a) (b) (c)

Fig. 2: Examples of stylized images: (a) the original image; (b) style transfer results using artistic styles; (c) style transfer result of our approach and the selected style. The introduction of color theme proximity produces more indistinguishable style transfer, compared with the use of artistic styles.

1) Style Selection Based on Color Proximity: We first quantize pixel values before calculating the color proximity.

Median Cut (MC). MC resembles a naive recursive clustering method that computes centroids of pixel values which will next serve as color themes for calculating color proximity [67]. In practice, the first frame of a clean video is selected to produce the color themes of the video. Since the target model classifies short videos filming single human actions, the first frames of the video could be visually similar to other frames. This selection scheme is proved effective through our experiments. The medians are used as the boundary to ensure that the pixels are divided into two subsets of the same size and can be separated further. Concretely, given an image z of n pixels, we let it be a set $z := \{z^R, z^G, z^B\}$ of pixel sets grouped by the RGB channels. Note that $z^c \subset \mathbb{R}^+,\ c \in \{R, G, B\}$, where z^c is the set of pixels in channel c. $\underline{z^c} = \inf\ z^c$ and $\overline{z^c} = \sup\ z^c$ are the infimum and the supremum of z^c, respectively. To obtain m color themes, first, 1) the pixels in z are sorted in ascending order of their pixel values in a channel c satisfying $c = \arg\max |\overline{z^c} - \underline{z^c}|$. Next, 2) the pixels are separated into two half sets $z_1 = \{z_{1,1}, ...z_{1,n/2}\},\ z_1 \leq median(z^c)$ and $z_2 = \{z_{2,(n+1)/2}, ...z_{2,n}\},\ z_2 > median(z^c)$ based on the median value $median(z^c)$. MC repeats 1) and 2) on z_1 and z_2 and their half sets till m sets of pixels are generated. Finally, the medians of the m pixel sets are returned as m quantized color themes of z.

Color Proximity. We define the color proximity between a

video and its style image as their Euclidean distance of color themes. A smaller proximity indicates a better style image with which style transfer can induce less perceptible changes in the original videos. In lieu of calculating the proximity in the RGB space, we turn to the HSV representation which aligns better with human perception [68]. We first project the RGB values obtained from MC into the HSV space. Next, the color proximity is measured in the HSV-projected XYZ space (see detailed HSV-to-XYZ transformation in Appendix D). Given the XYZ coordinate ϕ_i^x of the i-th color theme in the clean video x and ϕ_j^s of the j-th color theme in the style image s, the *one-versus-one* color proximity of s with respect to x is expressed as

$$d_{i,j}^s = \left\| \phi_i^x - \phi_j^s \right\|_2. \tag{2}$$

Furthermore, it should be noticed that the number of color themes affects the indistinguishability of the output videos. An insufficient number of color themes may lead to generating videos with unnatural content. This concern is emphasized in videos having a dominant but semantically less important background since the color themes are dominated by the background in most images. Therefore, as shown in Equation 3, we consider the C-*versus*-C color proximity between x and s. The style image s^* with the minimal C-*versus*-C color proximity is selected as the style image in style transfer. That is

$$s^* = \underset{s \in S_t}{\arg\min} \left(\sum_{i=1,j=1}^{C} d_{i,j}^s \right), \tag{3}$$

where s^* is the selected style image. In practice, we choose $C = 3$, as involving too many color themes enforces the style of the generated video taking after that of the style image. For untargeted attacks, the style image is selected based on Equation 3.

2) Stylized Boundary Search: To reduce the number of queries in targeted attacks, StyleFool conducts an offline search of style images that help approximate classification boundaries prior to sending queries to the target model. The fitness of each style image is evaluated by its prediction confidence in the target class. Due to the fact that style transfer minimizes the style loss between a source video and the style image, the stylized video is similar to the style image in their feature representations. When there exists a transferability between the feature extraction network and the target model, the video stylized by a style image in the target class is likely to be closer to the decision boundary of the target class. Therefore, adversarial videos can be crafted under fewer optimization steps after the stylization.

As presented in Equation 4, to select the best style image from the target class, we first convert each candidate image $s \in S_t$ in the style set S_t into the video x_{tile} by tiling s (*i.e.,* using s as the video frames repeatedly) according to the input frame requirement of the video classifier. The current top-1 label and the corresponding confidence score of $f(\cdot)$ on x_{tile} are obtained. If the top-1 label of x_{tile} is the target class y_t, we record its confidence score as $p(y_t|x_{tile})$. Otherwise, the confidence score of s will be set as 0.

$$score_s^t = \begin{cases} p\left(y_t \left| x_{tile} \right.\right), & if \ f\left(x_{tile}\right) = y_t, \\ 0, & otherwise. \end{cases} \tag{4}$$

In targeted attacks, we consider both the color theme proximity and $score_s^t$ of the style image at the same time. The principle of style image selection is updated as Equation 5 in targeted attacks.

$$s^* = \underset{s \in S_t}{\arg\min} \left(\sum_{i=1,j=1}^{C} d_{i,j}^s + \mu \left(1 - score_s^t\right) \right), \tag{5}$$

where μ is a weight coefficient.

Note that the color themes of all images and target class confidence of all targets can be calculated and stored before style transfer, which can save both query number and time cost when performing a batch video attack. In contrast, one of the state-of-the-art attacks, H-Opt [21], needs a significant amount of queries before the adversarial generation stage.

B. Style Transfer

With the potential defects of video style transfer in mind (such as the temporal inconsistency [63] and the flicker [64]), during the video style transfer stage, both stylized characteristics and temporal consistency should be considered [65]. The total loss of a video transfer can be expressed as

$$\begin{aligned} L_{total} = & \sum_{i=1}^{T} \left(\alpha L_{content}\left(x_i, x_i^s\right) + \beta L_{style}\left(x_i^s, s\right) + \gamma L_{tv}\left(x_i^s\right) \right) \\ & + \lambda \sum_{i=1}^{T-1} \left(L_{temporal}\left(x_i^s, x_{i+1}^s\right) \right), \end{aligned} \tag{6}$$

where s is the style image (to simplify the notation, we use s to present the style image in the rest of this paper); x_i is the i-th frame of the initial clean video; and x_i^s is the i-th frame of the stylized video x^s. $L_{content}$, L_{style} and L_{tv} represent the content loss, style loss, and total variation regularizer loss, respectively. α, β, γ, and λ are weight coefficients, and $L_{temporal}$ denotes the temporal loss between two consecutive frames.

Content loss. Content loss is presented as the normalized Euclidean distance between the initial video x and the stylized video x^s across all layers of the feature map in VGG-19 [69], which indicates the dissimilarity between the initial video and the stylized video in their high-level representations. Minimizing such content loss encourages x and x^s to share semantic similarities in their content. The content loss can be expressed as follows:

$$L_{content}\left(x_i, x_i^s\right) = \sum_k \frac{1}{H_k W_k C_k} \left\| \vartheta_k\left(x_i\right) - \vartheta_k\left(x_i^s\right) \right\|_2^2, \tag{7}$$

where H_k, W_k, and C_k represent the height, width and channel number of the k^{th} layer of the feature map, respectively. $\vartheta_k(\cdot)$ represents the feature map in the k^{th} layer.

Style loss. In order to reduce the difference between the style image and the stylized video, a style loss is introduced as the

sum of the style errors across all layers in the style transfer network, which could be expressed as follows:

$$L_{style}\left(x_i^s, s\right) = \sum_k \frac{1}{C_k^2} \left\| G_k\left(s\right) - G_k\left(x_i^s\right) \right\|_2^2, \tag{8}$$

where C_k is the number of channels in the k^{th} layer; $G_k\left(\cdot\right)$ represents the Gram matrix in the k^{th} layer [62]. The content loss and the style loss can move the input video towards or even across the decision boundary and alleviate the attack difficulty (*e.g.,* increase the attack success rate under query limit and reduce the query number) in black-box attacks, which are validated by the experiments in Section IV-B.

Total variation loss. A total variance loss can be regarded as a regular term in the loss function, which eliminates noise or shadow and improves the smoothness of each stylized image. Given a frame x_i^s, a sum of pairwise variance is computed over all pixels in x_i^s as follows:

$$\begin{aligned} L_{tv}\left(x_i^s\right) = \sum_{u,v} \Big(& \left\| x_i^s\left(u,v\right) - x_i^s\left(u+1,v\right) \right\|^2 \\ & + \left\| x_i^s\left(u,v\right) - x_i^s\left(u,v+1\right) \right\|^2 \Big), \end{aligned} \tag{9}$$

where $x_i^s\left(u,v\right)$ represents the pixel value at the coordinate (u,v) in x_i^s.

Temporal loss. The temporal loss is used to improve the consistency between two consecutive frames, which can be expressed as

$$L_{temporal}\left(x_i^s, x_{i+1}^s\right) = \frac{1}{HWC} O_{i+1} \left\| x_{i+1}^s - \mathcal{W}\left(x_i^s\right) \right\|^2, \tag{10}$$

where $\mathcal{W}\left(\cdot\right)$ represents a warp function [65] that outputs the warped frame generated from the input frame using pre-computed optical flow; O_{i+1} represents an occlusion mask matrix of the $\left(i+1\right)^{th}$ frame in optical flow.

Thus, the stylized video x^s is computed by minimizing the total loss in Equation 6, which has the characteristics of smoothness and high temporal consistency in parallel. Such temporal consistency in the stylized video can be retained in generated adversarial samples. Therefore, the introduction of temporal loss provides a guarantee for StyleFool to bypass the adversarial defense mechanism based on time consistency, *e.g.,* AdvIT [12].

Style loss and Maximum Mean Discrepancy. Another natural method would be optimizing the videos via a surrogate classifier prior to querying the target classifier. However, we discover that style transfer is superior to surrogate classifiers. It is known that minimizing the style loss is equivalent to minimizing the Polynomial kernel Maximum Mean Discrepancy (PMMD) between two sets $V_k\left(x_i^s\right)$ and $V_k\left(s\right)$ of feature vectors [70]. Let $M_k = H_k \times W_k$, $V_k\left(\cdot\right) := \left\{ \vartheta_k\left(\cdot\right)_p \left| \vartheta_k\left(\cdot\right)_p \in \mathbb{R}^{C_k} \right. \right\}_{p=1}^{M_k}$. $\vartheta_k\left(\cdot\right)_p$ is the p-th vector in the $C_k \times M_k$ feature map. Minimizing PMMD aligns the feature distributions of x_i^s and s. Such alignment may intrinsically increase the transferability of the optimized x^s on different networks since the frames of x^s are distributionally similar to s in the feature space. On the other hand, a surrogate classifier only minimizes the difference between the prediction of the optimized video and the target label without deliberately modifying the intrinsic feature distribution. Henceforth, the transferability of the surrogate-optimized video is susceptible to changes in the classifier parameters or architectures. As an evidence, the query numbers of attacks based on stylized videos are significantly less than those based on surrogate-optimized videos (check Table XII in Appendix C).

C. Adversarial Sample Generation

In the black-box attacks, since the attacker cannot have access to the model structure and parameters, it is impossible to establish the loss function and optimize it with the gradient information. In order to optimize the loss function in the black-box setting, Natural Evolution Strategy (NES), one of the most efficient gradient estimation methods [35] is introduced to estimate the gradients.

For an adversarial loss function $L\left(\theta\right)$ under a search distribution $\pi\left(\theta\left|x\right.\right)$, the gradient estimates $\hat{g}$ can be expressed as $\hat{g} = \nabla_x \mathrm{E}_{\pi(\theta|x)}\left[L\left(\theta\right)\right]$. By applying a log-trick inside the integral form of the above expectation, it yields

$$\begin{aligned} \nabla_x \mathrm{E}_{\pi(\theta|x)}\left[L\left(\theta\right)\right] &= \nabla_x \int L(\theta)\pi\left(\theta\left|x\right.\right) d\theta \\ &= \int L(\theta) \frac{\pi\left(\theta\left|x\right.\right)}{\pi\left(\theta\left|x\right.\right)} \nabla_x \pi\left(\theta\left|x\right.\right) d\theta \\ &= \mathrm{E}_{\pi(\theta|x)}\left[L\left(\theta\right) \nabla_x \log \pi\left(\theta\left|x\right.\right)\right]. \end{aligned} \tag{11}$$

In this way, it only requires to query the target model for the value of $L(\theta)$. To reduce the variance of the gradient estimates $\hat{g}$, we apply antithetic sampling to retrieve n points around x under a Gaussian search distribution, *i.e.,* half of the noise samples are generated by $\theta_i = x + \sigma\delta_i$ for $i \in \left\{1, ..., \frac{n}{2}\right\}$, where σ is the standard deviation of the noise, $\delta_i \sim N\left(0, I\right)$, I is an identity matrix. Then we invert the first half of the noise to obtain another half of the noise, *i.e.,* $\delta_i = -\delta_{n+1-i}$, for $i \in \left\{\frac{n}{2}+1, ..., n\right\}$. The gradient estimates can be approximated as follows:

$$\begin{aligned} \hat{g} \approx & \frac{1}{n\sigma} \sum_{i=1}^{n/2} \delta_i L\left(x + \sigma\delta_i\right) \\ & + \frac{1}{n\sigma} \sum_{i=n/2+1}^{n} \left(-\delta_{n+1-i}\right) L\left(x - \sigma\delta_{n+1-i}\right). \end{aligned} \tag{12}$$

Since antithetic sampling reduces the variance of the gradient estimates, the above equation can be regarded as a variance-reduced approximation of Stein's Lemma [71].

For black-box targeted attacks, only the top-1 label and its confidence score can be accessed, making it difficult to find the appropriate gradient descent direction [35], [66]. Therefore, the attack begins with an instance x_{adv} selected from the target class y_t and is optimized by Projected Gradient Decent (PGD) [72] with the gradient estimate. Since StyleFool has selected the most appropriate style image from the target class in the style transfer stage, the stylized video has carried a large number of features contained in the style image. In other words, higher similarity indicates closer distance

Algorithm 1: StyleFool.

Input: Black-box classifier f, input video x_0, input label y_0, target class y_t, style set S, perturbation threshold ε_{adv}, initial perturbation ε, total loss L_{total}, adversarial loss L.
Output: Adversarial video x_{adv}.

1 $S \leftarrow$ `init_style_set()`;
2 **for** $k \leftarrow 1$ *to* length(S) **do**
3 $c[k] \leftarrow$ `color_prox`$(x_0, S[k])$;
4 **if** $y_t \neq$ None **then**
5 $c[k] \leftarrow c[k] +$ `target_score`$(f, S[k], y_t)$;
6 $idx \leftarrow \arg\min(c)$;
7 $s^* \leftarrow S[idx]$;
8 $x^s \leftarrow$ `style_transfer`(x_0, s^*, L_{total});
9 **if** $y_t ==$ None **then**
10 $x_{adv} \leftarrow x^s$;
11 **while** $f(x_{adv}) == y_0$ **do**
12 $\hat{g} \leftarrow$ `NES`(x_{adv}, L);
13 $x_{adv} \leftarrow$ `PGD`$(x_{adv}, x^s, \hat{g}, \eta, \varepsilon_{adv})$;
14 **else**
15 $x_{adv} \leftarrow$ `to_video`(s^*);
16 **while** $\varepsilon > \varepsilon_{adv}$ *or* $f(x_{adv}) \neq y_t$ **do**
17 $\hat{g} \leftarrow$ `NES`(x_{adv}, L);
18 $x_{adv}, \eta, \varepsilon \leftarrow$ `BPGD`$(x_{adv}, x^s, \hat{g}, \eta, \varepsilon)$;

in high-dimensional space, which is beneficial to adversarial attacks. Therefore, we choose the video x_{vs} where the style image is located as the initial target class instance. In each PGD iteration, backtracking line search is introduced to find the minimal possible perturbation size which ensures that the target class y_t remains the top-1 label. Attack succeeds when the initial perturbation ε (initialized as 1) goes down and reaches the perturbation threshold ε_{adv}. This ensures the adversarial video generated from the stylized video is difficult to be distinguished by human eyes. In order to improve the attack efficiency, untargeted attacks directly take the stylized video x^s as the initial value of x_{adv}.

The whole process of StyleFool proposed in this paper is shown in Algorithm 1. In this algorithm, y_t is $None$ in the untargeted attack. `init_style_set` returns the initial style set which contains a large number of style images, `color_prox` outputs the color theme proximity, `target_score` gives out a target class confidence score of a style image, `style_transfer` delivers the style transfer, `to_video` finds the video to which the style image belongs, and `BPGD`$(\cdot)$ is an extension to PGD with backtracking line search and learning rate adjustment, the details of which are given in the work done by Ilyas *et al.* [35].

D. StyleFool Recap

In untargeted attack scenarios, for an input video, the style is selected based on the color theme proximity (line 3); while in targeted attack scenarios, the attacker will first randomly select a target label, then choose a style image in the target label according to the color theme proximity (line 3) and the target class confidence (line 5). Then, the input video is transferred to a stylized video according to the selected style (lines 6 to 8). Finally, NES and PGD are used to generate adversarial videos under black-box setting (lines 9 to 18). The required adversarial sample x_{adv} is generated at the end.

IV. Evaluation

In this section, we carry out comprehensive experiments to evaluate the performance and indistinguishability of StyleFool. Moreover, we provide possible variants of StyleFool, verify the importance of style selection in the ablation study, and reveal the advantage of style transfer through a quantitative evaluation.

A. Experiment Setup

Datasets. We employ two widely used datasets, UCF-101 [73] and HMDB-51 [74] to validate the attack performance of the proposed StyleFool.

- **UCF-101** is an action recognition dataset collected from YouTube, containing 13,320 video samples with 101 action classes, *e.g.,* archery, haircut, and punch.
- **HMDB-51** is a collection of realistic videos from various sources, including movies and web videos, containing 6,849 video samples with 51 action classes, *e.g.,* sword, climb, and golf.

We randomly select 70% of the videos in the datasets as the style set, and then the style image is selected from the style set using the criterion in Section III-A. The remaining 30% of the videos are used for adversarial sample generation.

Target video classifiers and evaluation metrics. We involve C3D [45] and I3D [24] as our targets. The two models utilize different strategies and achieve the state-of-the-art video classification performance. C3D learns both spatial and temporal features of input videos using 3D convolution, while I3D utilizes optical flow to build the relationship between two adjacent frames. Since C3D requires videos with 16 frames as input, we separate all videos into 16-frame snippets. The classification performance of C3D and I3D is shown in Table II. We define the following metrics in the evaluation.

- **Attack Success Rate (ASR)**: the ratio of adversarial videos that successfully mislead the classifier. Note that any attack exceeding the query limit will be considered as failed.
- **Minimal Queries (minQ), Maximal Queries (maxQ), and Average Queries (AQ)**: the minimal, maximal, and average numbers of queries to succeed in an attack. For fair comparison, in StyleFool, the number of queries during style selection is also counted, although this proportion is quite small (as analyzed in Section IV-B).
- **Indistinguishability**: the *naturalness* (*i.e.,* the generated adversarial videos are expected to preserve the semantic information of original videos and appear natural to human subjects) and *realness* (*i.e.,* the style transferred samples should retain the video quality of experience and mislead human subjects into thinking they are non-artificial videos) of videos. Additionally, the SSIM [75], PSNR, and FID [76] are used to statistically measure the indistinguishability, which are widely used to evaluate image quality. We extend them to video domain by averaging the per-frame metrics over the entire video. The equation of the PSNR is slightly

TABLE II: Video classification accuracy of C3D and I3D.

Model	Datasets	
	UCF-101	HMDB-51
C3D	84.9%	67.1%
I3D	87.6%	62.5%

modified to $PSNR = 10\log_{10}\frac{255^2}{(MSE+\alpha_m)}$ to avoid images with no difference, where MSE is the mean square error between two images, $\alpha_m = 10^{-5}$.

Benchmarking attack frameworks. In order to compare ℓ_p-norm attacks and unrestricted attacks, we extend the NES-based ℓ_p-norm attack on images (denoted by the first author's name as Ilyas [35]) and apply it onto video samples. Additionally, we compare our framework with two state-of-the-art one-on-one black-box attack frameworks, V-BAD [20] and H-Opt (Heuristic Opt-based Black-box Attack) [21]. V-BAD reduces the computational complexity by splitting tentative perturbations into several patches and rectifying the original gradient direction. Based on the Opt attack [36] in images, H-Opt uses Zeroth Order Optimization (ZOO) [34] to update the video. H-Opt considers adding sparse perturbations by key frames extraction and saliency detection.

Since C-DUP [18] and U3D [19] are universal perturbations only suitable for untargeted attacks, it is unreasonable and unfair to compare them with one-on-one attack frameworks (*e.g.,* V-BAD, H-Opt, and StyleFool) due to different attack purposes. Therefore, we only consider Ilyas [35], V-BAD [20], and H-Opt [21] as our benchmarks in the experiments. Considering that existing black-box attacks [32], [34], [35] require about 10^4 queries on CIFAR-10, and about 10^5 queries on ImageNet to succeed, we set a reasonable query limit as 3×10^5, which is similar to the setting of the research by Jiang *et al.* [20]. The perturbation threshold ε_{adv} is set as 0.05. We provide other configurations of StyleFool in Appendix A.

B. Experimental Results

Attack performance. In order to quantitatively compare the attack performance of StyleFool with benchmark attacks, we randomly select 100 videos from each dataset as the initial videos and mount both targeted and untargeted attacks on datasets UCF-101 and HMDB-51. Table III shows the attack performance of the four frameworks, *i.e.,* Ilyas, V-BAD, H-Opt, and StyleFool. Due to the restrictions on the number of queries (at most 3×10^5 queries) and partial information (the adversary can only access the top-1 label and its confidence score), the attack success rates of Ilyas, V-BAD and H-Opt cannot always achieve 100%, while StyleFool is stable at 100% for all scenarios; for some video samples, Ilyas, V-BAD and H-Opt cannot complete the attack within the query limit. Note that black-box query-based attacks will always succeed if there is no query limit. However, in StyleFool, even input videos that are far from the target class decision boundary will not cause the query number to exceed the upper limit, as style transfer will push the adversarial samples towards the boundary and thus fewer queries are ensured at the same time.

According to the average queries results, compared with the benchmark frameworks, StyleFool requires the least number of queries to succeed in an attack. Specifically, the query number of StyleFool is at least 69% and 44% and 72% less than Ilyas, V-BAD and H-Opt, respectively, in targeted attacks on C3D using the UCF-101 dataset. When the attacked model is I3D, in untargeted attacks conducted on the HMDB-51 dataset, the average queries of StyleFool are far fewer than those of H-Opt (2,013 vs. 37,897). On average, StyleFool reduces 65% (77%), 43% (53%), and 83% (87%) of queries to succeed in the targeted (untargeted) attack, compared with Ilyas, V-BAD and H-Opt, respectively. Overall, StyleFool requires the least queries to succeed in an attack. Specifically, in cases where the number of queries is 1 (*e.g.,* the minQs in untargeted attacks), the stylized video is already adversarial, and thus no more queries are needed. As a result, StyleFool demonstrates the efficiency of attacking with $(2.4 \sim 6.5) \times 10^4$ queries for targeted attacks and $(1.5 \sim 6.6) \times 10^3$ queries for untargeted attacks. The experimental results demonstrate that StyleFool is more generalized and efficient, compared with the state-of-the-art approaches. Figure 5 in Appendix C shows the adversarial video samples generated by StyleFool in targeted and untargeted attacks. The styles of original videos have been transferred and the goal of adversarial attack has also been achieved.

Analysis. The reason why Ilyas requires more queries is that in the high-dimensional feature space, it is more difficult to pull the video from the target class back into the ℓ_p ball of the original video, and randomly selecting initial target class video makes the attack difficulty uncertain. Note that Ilyas is a special case of StyleFool without style transfer. Since H-Opt needs many queries in finding the initial attack direction and calculating every decision boundary distance using fine-grained search and binary search, the average queries remain high as a whole. When the query limit is set, the attack fails when the number of queries is greater than 3×10^5. Therefore, the attack success rate of H-Opt is significantly lower than that of the other two frameworks. Though H-Opt considers sparse perturbations temporally and spatially, such a weak advantage cannot be offset by the high cost of query number which is one order of magnitude higher than other frameworks, since query number is the most important index of black-box attack. V-BAD improves more than H-Opt. However, its performance is still not as good as StyleFool. The style transfer in StyleFool moves videos closer to the decision boundary, which is beneficial to adversarial attacks, resulting in fewer queries. The average queries of StyleFool are 40-70% fewer than that of V-BAD.

Note that in the style selection stage, we need to query the classifier for targeted attacks to obtain the target class confidence of all style images, but the number of queries is extremely smaller in comparison with the adversarial sample generation stage. When attacking a C3D model on the UCF-101 dataset, H-Opt spends about 15% of queries (approximately 3×10^4) on searching an initial target class video and extracting key frames, while StyleFool spends only 0.45%

TABLE III: Attack performance comparison.

Model	Attack	UCF-101 (Targeted)				HMDB-51 (Targeted)				UCF-101 (Untargeted)				HMDB-51 (Untargeted)			
		ASR	minQ	maxQ	AQ	ASR	minQ	maxQ	AQ	ASR	minQ	maxQ	AQ	ASR	minQ	maxQ	AQ
C3D	Ilyas [35]	62	43,112	>300,000	>209,847	96	4,302	>300,000	>96,563	**100**	50	58,948	24,202	**100**	50	47,727	6,631
	V-BAD [20]	85	1,269	>300,000	>117,053	98	517	>300,000	>56,767	**100**	50	150,725	12,646	**100**	50	42,974	4,083
	H-Opt [21]	60	16,989	>300,000	>228,867	70	10,010	>300,000	>195,242	**100**	1,020	97,929	19,131	**100**	1,948	42,183	14,066
	StyleFool	**100**	**1,284**	**248,131**	**65,066**	**100**	**101**	**95,897**	**35,232**	**100**	**1**	**19,013**	**3,811**	**100**	**1**	**9,948**	**1,526**
I3D	Ilyas [35]	95	10,875	>300,000	>91,495	**100**	4,344	246,149	64,047	**100**	1,177	139,749	25,187	**100**	197	33,811	7,234
	V-BAD [20]	99	101	>300,000	>56,680	**100**	521	211,138	42,508	**100**	148	123,236	10,437	**100**	50	30,822	3,447
	H-Opt [21]	31	8,583	>300,000	>260,463	53	6,892	>300,000	>212,256	**100**	3,165	156,624	47,008	**100**	1,386	164,580	37,897
	StyleFool	**100**	**101**	**119,991**	**30,876**	**100**	**101**	**80,319**	**23,556**	**100**	**1**	**30,724**	**6,643**	**100**	**1**	**6,665**	**2,013**

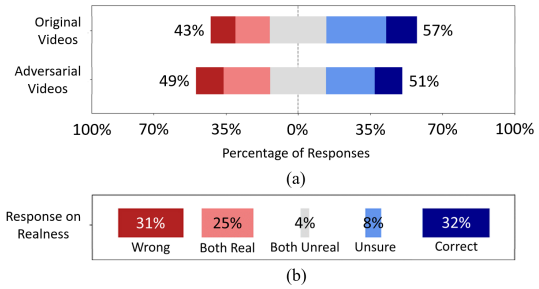


Fig. 3: Results of the user study on indistinguishability: (a) Participants' responses in the naturalness test. ■ : very unnatural, ■ : very natural; (b) Participants' responses in the realness test.

of queries (approximately 3×10^2) during the style selection stage.

C. Indistinguishability

In order to verify the indistinguishability of StyleFool, we conducted a user study to evaluate the naturalness and realness of adversarial samples (randomly selected from the samples generated from the experiment in Section IV-B). We designed a web questionnaire where both adversarial videos and clean videos are displayed and recruited 160 anonymous participants in an online survey via Amazon Mechanical Turk (AMT) [77]. The Human Research Ethics Committee of lead author's affiliation determined that the study was exempt from further human subjects review, and we followed best practice for ethical human subjects survey research, *e.g.,* all questions were optional and we did not collect unnecessary personal information. Participant consented for their answers to be used for academic research. All participants are over 18 years old and are able to complete the survey in English. Appendix B provides the survey protocol and more detailed settings (*e.g.,* the demographics and the payment).

Survey results and data analysis. The average completion time of participants in the naturalness test and the realness test was 59.4 minutes and 62.5 minutes respectively. According to the participants' responses about naturalness shown in Figure 3(a), we find that participants could not distinguish between clean videos and adversarial videos very well. 51% of participants gave positive naturalness responses to adversarial videos, while 57% of participants thought the original clean videos were natural – only 6% higher. It can be concluded that most participants did not give a high confidence score of naturalness to clean videos (15% of participants voted "very natural" for clean videos), nor did they give a low confidence of naturalness to the adversarial video (13% of participants voted "very unnatural" on adversarial samples). Our variables are ordinal and the responses to each question are not expected to be normally distributed. Therefore, we conduct a Mann-Whitney U-test [78] for statistical significance testing. Concretely, we calculate the p-value of the aggregated Likert scale responses on original videos and adversarial videos with the null hypothesis H_0: there is no significant difference between the naturalness of original videos and adversarial videos. The U-test result ($p = 0.07 > 0.05$) suggests that we cannot reject the null hypothesis, indicating a high degree of sensory comfort is maintained by adversarial videos generated by StyleFool.

Figure 3(b) shows the distribution of participants' responses in the realness test. Only 32% of the participants correctly distinguished adversarial videos from clean videos, and 31% of the participants provided totally incorrect answers. Considering that the accurate rate is much lower than random guessing, we can conclude that the adversarial samples could not be identified by participants. Additionally, 25% of participants said both videos were real, which indicates that adversarial videos looked as realistic as clean videos for them. 4% of the participants thought that both videos were unreal, and 8% got trapped in judgment difficulty. The accuracy of participants' judgment is 50.5% with 50.4% precision and 62.0% recall, which indicates that the participants could not distinguish the adversarial videos well.

Quantitative analysis. Since StyleFool is an attack with unrestricted perturbations, we argue that it could be unfair to compare unrestricted attacks and restricted attacks by video quality metrics. However, we still would like to quantitatively evaluate the indistinguishability performance of StyleFool. We report the SSIM, PSNR, and FID between original videos and adversarial videos, shown as "StyleFool-ori-adv" in Table IX in Appendix C, which is expected to be lower than existing restricted methods (such as V-BAD and H-Opt), but still good enough to keep stealthy. The changes in styles and textures are bound to cause differences in those three metrics on video quality, but the semantic information is not influenced (see Figure 5 in Appendix C). We further evaluate the indistinguishability between stylized videos and adversarial videos (*i.e.,* the "StyleFool-sty-adv" in Table IX). As we particularly restrict the perturbations between stylized videos and the adversarial videos, the "StyleFool-sty-adv" shows the best indistinguishability performance among all benchmark

methods, due to less modification in pixels.

D. Ablation Study

Possible variants of StyleFool. Since StyleFool is an unrestricted attack, the perturbation threshold, ε_{adv}, could be varied, instead of strictly setting to 0.05 in the process of PGD. Properly increasing ε_{adv} may not affect the visual sense. To determine the influence of larger perturbation, we set ε_{adv} to 0.05, 0.10, 0.15 and 0.20, respectively, and conduct ablation experiments on two datasets with two models. Table IV and Figure 6 in Appendix C show the attack performance and visualization of StyleFool under different ε_{adv}. The average queries are reduced sharply with the increase of ε_{adv}. However, since targeted attacks start with a video from the targeted label, a larger ε_{adv} will lead to the superposition of two videos in the adversarial video (*e.g.,* $\varepsilon_{adv} = 0.20$ in targeted attacks shown in Figure 6(a)), although the average queries are reduced by an order of magnitude. We point out that there is a trade-off between the number of queries and visual perception. While in untargeted attacks, a larger ε_{adv} will not exert much impact on the visual perception of the adversarial video (*e.g.,* $\varepsilon_{adv} = 0.10, 0.15$, and 0.20 in untargeted attacks shown in Figure 6(b)). In addition, StyleFool with a larger ε_{adv} can largely reduce queries, *e.g.,* in untargeted attacks against C3D trained on the HMDB51 dataset, StyleFool with $\varepsilon_{adv} = 0.20$ reduces average queries by 73%, compared with $\varepsilon_{adv} = 0.05$. To conclude, a larger perturbation threshold is practicable in StyleFool, but the trade-off between the number of queries and visual perception should also be considered, especially in targeted attack scenarios.

Contribution of style selection in StyleFool. To further explore how color proximity and target class confidence contribute to the performance of StyleFool, we next conduct an ablation study. Concretely, we carry out experiments in the following four different style selection scenarios: (*i*) randomly selecting styles; (*ii*) only color theme proximity is considered (StyleFool in untargeted attacks); (*iii*) only target class confidence is considered; (*iv*) both color theme proximity and target class confidence are considered (StyleFool in targeted attacks). We perform all 4 scenarios for targeted attacks, and settings (*i*) and (*ii*) for untargeted attacks. In each scenario, we randomly select 50 videos from UCF-101 and HMDB-51 (25 from each), and use C3D as the target model.

Table V shows the StyleFool experimental results of the ablation study. Figure 7 in Appendix C visualizes two video samples generated in targeted and untargeted scenarios. Note that the performance of StyleFool is reflected by not only the attack success rate and query number (Table V), but also the naturalness and realness (Figure 7). Only when all the metrics are satisfied, we consider an attack effective. Recall that StyleFool can reduce a large number of queries and thus ensure the attack success rate up to 100%. All the attack success rates in the ablation study are 100% since all attacks succeed within the query limit. In targeted scenarios, if the style is selected randomly, not only the adversarial video is easy to be distinguished (*e.g.,* weird colors and textures in Figure 7(b)), but also the query number is not reduced. Although color theme proximity brings higher similarity and indistinguishability in the stylized video, more queries are needed if we only consider the color theme proximity, since the output distribution of the classifier does not experience a major change after transfer. When only considering target class confidence, the number of queries can be reduced as stylized videos can be moved closer to the decision boundary. The indistinguishability, however, is not fairly satisfying (*e.g.,* the blue leaves in Figure 7(d)). Significantly, when the two criteria are both considered in style selection, the adversarial videos have high sensory comfort and the attack is efficient. In untargeted scenarios, similar to targeted attacks, stylized videos have higher sensory comfort when color theme proximity is considered (observed from Figures 7(g) and (h)). Note that the average queries are also guaranteed to be about 25% less than random selection.

Contribution of style transfer in StyleFool. In order to quantitatively evaluate the contribution of style selection and style transfer in StyleFool, we further compare the confidence scores before and after the style transfer process, *i.e.,* the original video in original (target) class vs. the stylized video in original (target) class.

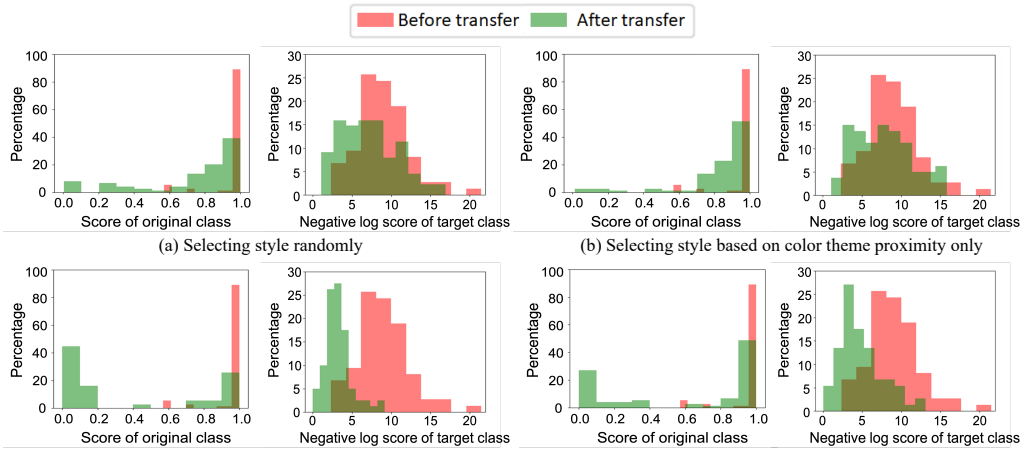
Figure 4 shows the confidence scores of samples in four style selection scenarios mentioned in the ablation study. According to the scores of original class on the left side, before transfer, most original samples have confidence scores of 1.0 in their original classes. The scores decrease significantly after style transfer, which means that style transfer can actively bring the video sample close to or even across the decision boundary. In order to show the change of target class scores after style transfer, we present the distribution in a negative log score manner. The decrease of negative log scores after the style transfer indicates that the stylized videos have been moved closer to the decision boundary of the target class. Even though the style transfer cannot make every stylized video classified as the target class (*i.e.,* make the target class as top-1 score), such a move toward decision boundary will lead to the reduction of queries in the subsequent adversarial generation and boost the efficiency of attack. It is worth noting that, in some cases, stylized video samples can mislead the classifiers and succeed in the untargeted attack even without the adversarial sample generation process, as the confidence score of original class has already been significantly reduced and is no longer predicted as the top-1 score. No further queries are needed in such cases, *e.g.,* over 50% of the stylized videos can successfully fool the C3D classifier under untargeted attacks for UCF-101, which provides a reasonable explanation that the minQ of StyleFool is 1 in Table III.

V. Countermeasures

In this section, we show the performance of StyleFool against state-of-the-art video defenses and further discuss other mitigation strategies.

TABLE IV: StyleFool performance with different ε_{adv}.

Model	ε_{adv}	UCF-101 (Targeted)				HMDB-51 (Targeted)				UCF-101 (Untargeted)				HMDB-51 (Untargeted)			
		ASR	minQ	maxQ	AQ	ASR	minQ	maxQ	AQ	ASR	minQ	maxQ	AQ	ASR	minQ	maxQ	AQ
C3D	0.05	100	1,284	248,131	65,066	100	101	95,897	35,232	100	1	19,013	3,811	100	1	9,948	1,526
	0.10	100	101	127,432	25,054	100	101	62,546	15,240	100	1	7,204	1,348	100	1	2,842	620
	0.15	100	101	83,134	12,297	100	101	33,268	6,490	100	1	5,538	1,155	100	1	2,402	457
	0.20	100	101	57,223	6,553	100	101	16,227	2,202	100	1	3,803	813	100	1	2,059	415
I3D	0.05	100	101	119,991	30,876	100	101	80,319	23,556	100	1	30,724	6,643	100	1	6,665	2,013
	0.10	100	101	82,588	10,502	100	101	51,478	10,378	100	1	18,131	3,145	100	1	2,892	1,007
	0.15	100	101	63,918	5,050	100	101	29,953	4,653	100	1	16,220	2,802	100	1	2,304	927
	0.20	100	101	40,895	2,237	100	101	20,344	2,480	100	1	13,280	1,873	100	1	2,108	899



(a) Selecting style randomly

(b) Selecting style based on color theme proximity only

(c) Selecting style based on confidence score only

(d) Both (targeted StyleFool)

Fig. 4: Confidence scores of the original class and target class before and after style transfer with different style selection strategies.

TABLE V: Results of StyleFool in different scenarios: (*i*) randomly selecting styles; (*ii*) only color theme proximity is considered (untargeted StyleFool); (*iii*) only target class confidence is considered; and (*iv*) both color theme and target class confidence are considered (targeted StyleFool).

Attack	UCF-101 (Targeted)				HMDB-51 (Targeted)			
	ASR	minQ	maxQ	AQ	ASR	minQ	maxQ	AQ
Random	**100**	**101**	191,600	52,285	**100**	**101**	154,299	37,232
Color theme proximity	**100**	7,416	257,676	77,642	**100**	5202	149,232	49,227
Target class confidence	**100**	**101**	211,110	55,534	**100**	**101**	156,875	43,575
Both	**100**	**101**	**106,713**	**44,096**	**100**	**101**	**68,850**	**31,894**
Attack	**UCF-101 (Untargeted)**				**HMDB-51 (Untargeted)**			
	ASR	minQ	maxQ	AQ	ASR	minQ	maxQ	AQ
Random	**100**	**1**	18,416	2,837	**100**	**1**	11,712	1,236
Color theme proximity	**100**	**1**	**13,563**	**2,228**	**100**	**1**	**4,068**	**937**

A. Performance Against Defenses

Defense algorithms. In order to further reflect the performance of StyleFool, we select three state-of-the-art video adversarial defenses and one detection method to evaluate the adversarial samples.

- **AdvIT** [12] is the first video defense method, which uses optical flow to generate pseudo frames and evaluate the classifier output consistency between the pseudo frame and the target frame. As it is an adversarial frame detection method that only detects whether a video has been attacked, we use Area Under Curve (AUC) to measure its defense performance.
- **ComDefend** [13] is a spatial defense method for videos, which has been claimed to have better defense performance against dense attacks, compared with the temporal defense [14]. ComDefend employs a CNN (ComCNN) for compressing videos and another CNN (RecCNN) for video reconstruction. We use Defense Success Rate (DSR), the ratio of adversarial videos which are successfully defended, to evaluate the defense performance of ComDefend. Concretely, if the adversarial noise is successfully removed and the denoised video is correctly classified into the original class, the defense is considered as successful.
- **RS** [15] is a certified defense method in images, which utilizes noise sampled from Gaussian distributions to smooth the predicted scores and provides certifiably robust regions in ℓ_p norms to data samples [15], [79], [80]. RS has been proved to certify ℓ_2 perturbations, but is not effective for $\ell_p\,(p > 2)$, especially ℓ_∞ perturbations [81], [82]. We extend RS to videos and test the defense performance of video attacks. We use Clean Accuracy (CA), Adversarial Accuracy (AA), and Average Certified Radius (ACR) for defense evaluation. CA and AA indicate the prediction success rate of the smoothed classifier in the clean samples and adversarial samples, respectively. ACR represents the average value of the maximum certified radius of the smoothed classifier.
- **CNN-generated Image Detector** [43] is a detector for GAN-generated images, which trains a binary classifier to distinguish whether the image is generated or real. We extend CNN-generated image detection to videos to eval-

TABLE VI: AUC performance of AdvIT and DSR performance of ComDefend against adversarial attacks.

Model	Attack	AUC of AdvIT		DSR of ComDefend	
		UCF-101	HMDB-51	UCF-101	HMDB-51
C3D	Ilyas [35]	99%	97%	85%	88%
	V-BAD [20]	97%	94%	76%	78%
	H-Opt [21]	93%	92%	74%	75%
	StyleFool	**58%**	**54%**	**36%**	**34%**
I3D	Ilyas [35]	98%	95%	95%	93%
	V-BAD [20]	96%	93%	87%	78%
	H-Opt [21]	95%	93%	82%	75%
	StyleFool	**53%**	**52%**	**53%**	**48%**

uate the performance of StyleFool, since StyleFool can be considered as a generative approach to some extent.

In AdvIT, 50 adversarial videos are randomly selected from UCF-101 and 50 from HMDB-51. For AdvIT, 5 frames are selected from each adversarial video as target frames, and the average KL divergence is calculated as the consistency using optical flow between the target frame and its previous three frames. Columns 3 and 4 in Table VI show the AUC performance of AdvIT against the four attack frameworks. StyleFool outperforms the other three benchmark frameworks with the lowest defense AUC. Although AdvIT utilizes optical flow to evaluate the temporal consistency of the adversarial video, StyleFool compromises AdvIT via considering the temporal loss in style transfer, which can increase the temporal consistency of stylized videos. In addition, since fewer queries are needed in the adversarial sample generation stage, the temporal consistency will not lose too much. Thus, it is not surprising that StyleFool performs better against AdvIT.

Columns 5 and 6 in Table VI report the defense performance of ComDefend. In all scenarios, the defense success rates of StyleFool are lower than those of three benchmark frameworks, which indicates that ComDefend cannot turn adversarial videos from StyleFool back to clean videos. Specifically, when the target model is C3D, the defense success rates of Ilyas, V-BAD and H-Opt under two datasets are all higher than 74%, while the defense success rates against StyleFool are less than 36% (nearly half of 74%). Even when attacking I3D, the defense success rates of StyleFool (both around 50%) are still close to blind guess. Since ComDefend adds a Gaussian noise to the output of ComCNN, the adversarial perturbations can be removed after reconstructing the video in RecCNN. However, such a defense mechanism can only work for restricted perturbations. Since the mean of Gaussian noise added after ComCNN is 0, ComDefend is hard to defend against unrestricted perturbations.

In RS, we set the number of Monte Carlo samples used for estimation n as 10,000, and the variance of Gaussian noise σ as 0.25. The other parameters remain at their default values [15]. We argue that a larger n is more convincing in video scenarios, due to the high dimension of video samples. Table VII reports the defense performance of RS. In all scenarios, the clean accuracy exceeds 54% (slightly lower than 65% when applied in image scenarios [15]). Due to the consideration of unrestricted perturbations in StyleFool, the adversarial accuracy is less than 13%, and the average certified radius is less than 0.06, indicating that StyleFool can break the robustness certificate in a breeze. To conclude, adversarial examples crafted by StyleFool can escape from the robust regions induced by RS since StyleFool can conceal perturbations large in ℓ_p norms in unsuspicious patterns. To defend against StyleFool, RS is required to obtain larger robust regions with higher security budgets on the protected model, and this leads to a substantial utility trade-off, or even results in loss of utility.

TABLE VII: Performance of Randomized Smoothing against adversarial attacks.

Model	Attack	UCF-101			HMDB-51		
		CA	AA	ACR	CA	AA	ACR
C3D	Ilyas [35]	66%	52%	0.29	63%	50%	0.29
	V-BAD [20]		51%	0.29		49%	0.27
	H-Opt [21]		44%	0.22		34%	0.19
	StyleFool		**13%**	**0.06**		**7%**	**0.02**
I3D	Ilyas [35]	54%	46%	0.26	56%	46%	0.27
	V-BAD [20]		45%	0.27		45%	0.24
	H-Opt [21]		38%	0.21		41%	0.22
	StyleFool		**9%**	**0.04**		**7%**	**0.03**

TABLE VIII: Performance of CNN-generated image detection on StyleFool.

Model	Dataset	Real Acc	Fake Acc
C3D	UCF-101	99.74%	27.89%
	HMDB-51	99.38%	27.66%
I3D	UCF-101	99.76%	29.11%
	HMDB-51	99.50%	26.88%

Table VIII shows the experimental results of applying CNN-generated image detection on StyleFool. The method achieves near 100% detection accuracy on real video samples (Real Acc), but less than 30% detection accuracy on adversarial samples (Fake Acc). The reason is perhaps because StyleFool introduces unrestricted perturbations in adversarial examples. Such a detection method that focuses on finding GAN-generated images without perturbations is not suitable.

B. Other Mitigation Strategies

Adversarial training. As one of the classical defenses against adversarial attacks, Adversarial Training (AT) can improve the accuracy of the classifier by training a mixture of clean samples and adversarial samples [16], [17]. Adversarial training could be effective when the perturbations are restricted, as the possible perturbations can be effectively computed, resulting in a trained robust classifier. However, as mentioned previously, the unrestricted perturbations introduced by StyleFool have distinct distributions with that of restricted perturbations. Moreover, it is hard for the defender to determine possible perturbations since the perturbations can be diversified by choosing different style images. Therefore, the capability of adversarial training against StyleFool could be very limited.

Object detection with humans in the loop. With respect to the object detection in a video, although the output score of an object can be reduced much after StyleFool, human eyes still provide a high confidence for that object in the adversarial video. For example, a car in the adversarial sample

may not be detected by an object detector due to StyleFool, but human eyes will always recognize the car even if its style has been transferred. Therefore, the adversarial videos can be distinguished from clean videos by human sensory cognition when the labels are inconsistent with the objects in the video. However, such mitigation will cost much and is even impossible for large-scale detection.

VI. Discussion

Temporal consistency. We guarantee the temporal consistency of the stylized video in the style transfer stage. Although irregular perturbations are generated in the adversarial attack stage, the detection results under AdvIT show that adversarial perturbations have little impact on the temporal consistency of adversarial videos. We also evaluate the SSIM between stylized videos and adversarial videos, as shown in Table IX. We find that the SSIMs under targeted attacks are all greater than 0.76 and those under untargeted attacks are all greater than 0.80, which indicate that videos before and after attacks have high similarity. It is worth noting that if the temporal loss in style transfer is added to the loss function in the adversarial attack stage, the consistency of the adversarial video may be further improved, and AdvIT will find it more difficult to detect the adversarial videos. However, such a strategy may increase the number of queries, since it requires the perturbations between two adjacent frames to be related.

Online universal attacks. Although StyleFool focuses on one-on-one offline attacks, our work can further support universal online attacks through certain changes. Similar to [63], [64], we can train a style transfer model for each style. When the input is a batch of videos, we only need to select the style image that can ensure sensory comfort according to the color theme proximity, and then input the video into the model corresponding to the style image to obtain the stylized video. Such procedure is similar to the previous work [18], [19] where the universal perturbation is first trained offline and then superimposed on the online video. We carry out a pilot study on a small batch of videos, and find that around 75% of stylized videos can be directly misclassified (query is not allowed since it is online untargeted attacks), which is very close to the attack success rates of C-DUP and U3D (about 80%). We will further explore the capability of applying StyleFool in the online attack scenarios in the future.

Style selection for long videos. The state-of-the-art video classifiers involved in our research allow only 16-frame video inputs. The changes in the video frames with respect to colors and actions are small, as the duration of a 16-frame video is usually less than 1 second (played at 24 frames per second), which benefits the StyleFool quite a lot. We provide an analysis of the frame variation in Appendix E. For longer videos in the real world, the selection of the style image could be more complex, as in a long video, the action of the characters and the style of the scene may change frequently, and the performance of attack could be negatively influenced. A potential attack strategy could be extracting multiple style images at equal intervals and adding them to the style set as style images.

Style detection. Various AI technologies are used to classify styled images [83]–[86]. However, these style detectors can only classify pre-defined artistic styles, such as realism, romanticism, and symbolism. Considering that the styles involved in our study are all from real life with numerous (or even countless) styles, it is impossible to learn the style of each image through training, *e.g.,* mapping a specific video content to a specific style. On the other hand, if we input the original video (or a frame) and an adversarial video into a style detector, it cannot tell which style is the original style and which one is generated, although the videos might be classified as two different styles. Further, the adversarial perturbations will also make it difficult to distinguish whether a video is stylized, since a small perturbation can greatly change the detection results.

VII. Conclusion

This paper mainly studies the video black-box adversarial attack, and proposes StyleFool against video classification systems. StyleFool generates unrestricted perturbations without changing semantic information, jumping out of the shackles of traditional restricted attacks. StyleFool selects style images based on the color theme proximity and the target class confidence, and transfers the initial video with the best style. Finally, NES is used for gradient estimation to solve the black-box setting. Experimental results show that the stylized videos can change the video style without changing the semantic content of the video, as well as push videos closer to or even across the decision boundary. StyleFool performs the best in both targeted and untargeted attacks. We also show that the adversarial videos generated by StyleFool are indistinguishable to human eyes, and the perturbations are observable but imperceptible. Finally, StyleFool can also evade the existing video defenses since most defenses are tailored for restricted perturbations. As a byproduct, we illustrate that the human-centric metrics developed in this paper can not only make stylized videos look natural to human eyes, but also make target class scores increase sharply, reducing the query number in adversarial attacks dramatically. In future work, we hope to explore possible new defenses against StyleFool.

Acknowledgment

The authors are grateful to the anonymous reviewers for their feedback that helped improve the paper. This work was supported in part by the National Natural Science Foundation of China (61972219), the Research and Development Program of Shenzhen (JCYJ20190813174403598, SGDX20190918101201696, 20210324120012033), the Overseas Research Cooperation Fund of Tsinghua Shenzhen International Graduate School (HW2021013), Guangdong Provincial Key Laboratory of Cyber and Information Security Vulnerability Research (No.2020B1212060081). Xi Xiao is the corresponding author of this paper.

REFERENCES

[1] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[2] Q. Kong, M.-A. Rizoiu, S. Wu, and L. Xie, "Will this video go viral: Explaining and predicting the popularity of youtube videos," in *Companion Proceedings of the The Web Conference (WWW)*, 2018, pp. 175–178.

[3] X. Cheng, J. Liu, and C. Dale, "Understanding the characteristics of internet short video sharing: A youtube-based measurement study," *IEEE Transactions on Multimedia*, vol. 15, no. 5, 2013.

[4] "Cisco annual internet report (2018–2023) white paper," https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

[5] B. Liu, M. Wu, M. Tao, Q. Wang, L. He, G. Shen, K. Chen, and J. Yan, "Video content analysis for compliance audit in finance and security industry," *IEEE Access*, vol. 8, pp. 117 888–117 899, 2020.

[6] M. Marciel, R. Cuevas, A. Banchs, R. González, S. Traverso, M. Ahmed, and A. Azcorra, "Understanding the detection of view fraud in video content portals," in *Proceedings of the 25th International Conference on World Wide Web (WWW)*, 2016.

[7] K. Yuan, D. Tang, X. Liao, X. Wang, X. Feng, Y. Chen, M. Sun, H. Lu, and K. Zhang, "Stealthy porn: Understanding real-world adversarial images for illicit online promotion," in *Proceedings of the IEEE Symposium on Security & Privacy (SP)*, 2019, pp. 952–966.

[8] "Facebook publishing child pornography," https://www.thetimes.co.uk/article/facebook-publishing-child-pornography-pdgt87nm6.

[9] "Deepfakes github," https://github.com/deepfakes/faceswap.

[10] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2387–2395.

[11] J. Wakefield. (2022) Deepfake presidents used in Russia-Ukraine war. https://www.bbc.com/news/technology-60780142.

[12] C. Xiao, R. Deng, B. Li, T. Lee, B. Edwards, J. Yi, D. Song, M. Liu, and I. Molloy, "Advit: Adversarial frames identifier based on temporal consistency in videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3968–3977.

[13] X. Jia, X. Wei, X. Cao, and H. Foroosh, "Comdefend: An efficient image compression model to defend adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6084–6092.

[14] X. Jia, X. Wei, and X. Cao, "Identifying and resisting adversarial videos using temporal consistency," *arXiv preprint arXiv:1909.04837*, 2019.

[15] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019, pp. 1310–1320.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[17] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[18] S. Li, A. Neupane, S. Paul, C. Song, S. V. Krishnamurthy, A. K. Roy-Chowdhury, and A. Swami, "Stealthy adversarial perturbations against real-time video classification systems." in *Proceedings of the Symposium on Network and Distributed Systems Security (NDSS)*, 2019.

[19] S. Xie, H. Wang, Y. Kong, and Y. Hong, "Universal 3-dimensional perturbations for black-box attacks on video recognition systems," in *Proceedings of the IEEE Symposium on Security & Privacy (SP)*, 2022.

[20] L. Jiang, X. Ma, S. Chen, J. Bailey, and Y.-G. Jiang, "Black-box adversarial attacks on video recognition models," in *Proceedings of the 27th ACM International Conference on Multimedia (ACM MM)*, 2019, pp. 864–872.

[21] Z. Wei, J. Chen, X. Wei, L. Jiang, T.-S. Chua, F. Zhou, and Y.-G. Jiang, "Heuristic black-box adversarial attacks on video recognition models," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020, pp. 12 338–12 345.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1097–1105, 2012.

[23] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 1, pp. 221–231, 2012.

[24] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.

[25] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *Proceedings of 25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 513–530.

[26] S. Chen, C. Peng, L. Cai, and L. Guo, "A deep neural network model for target-based sentiment analysis," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–7.

[27] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[28] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of the IEEE Symposium on Security & Privacy (SP)*, 2017, pp. 39–57.

[29] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation (TEC)*, vol. 23, no. 5, pp. 828–841, 2019.

[30] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2574–2582.

[31] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016, pp. 372–387.

[32] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (AsiaCCS)*, 2017, pp. 506–519.

[33] A. N. Bhagoji, W. He, B. Li, and D. Song, "Practical black-box attacks on deep neural networks using efficient query mechanisms," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 154–169.

[34] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 15–26.

[35] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 2137–2146.

[36] M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," *arXiv preprint arXiv:1807.04457*, 2018.

[37] M. Zajac, K. Zołna, N. Rostamzadeh, and P. O. Pinheiro, "Adversarial framing for image and video classification," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 10 077–10 078.

[38] X. Wei, J. Zhu, S. Yuan, and H. Su, "Sparse adversarial perturbations for videos," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 8973–8980.

[39] S. Li, A. Aich, S. Zhu, S. Asif, C. Song, A. Roy-Chowdhury, and S. Krishnamurthy, "Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 2085–2096, 2021.

[40] Z. Wang, C. Sha, and S. Yang, "Reinforcement learning based sparse black-box adversarial attack on video recognition models," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[41] J.-W. Chang, M. Javaheripi, S. Hidano, and F. Koushanfar, "Adversarial attacks on deep learning-based video compression and classification systems," *arXiv preprint arXiv:2203.10183*, 2022.

[42] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[43] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[44] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[45] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[46] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2625–2634.

[47] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4694–4702.

[48] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[49] H. Hosseini and R. Poovendran, "Semantic adversarial examples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018.

[50] A. Bhattad, M. J. Chong, K. Liang, B. Li, and D. A. Forsyth, "Unrestricted adversarial examples via semantic manipulation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[51] A. S. Shamsabadi, R. Sanchez-Matilla, and A. Cavallaro, "Colorfool: Semantic adversarial colorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1151–1160.

[52] C. Yang, A. Kortylewski, C. Xie, Y. Cao, and A. Yuille, "Patchattack: A black-box texture-based attack with reinforcement learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[53] R. Duan, X. Ma, Y. Wang, and J. Bailey, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1000–1008.

[54] T. B. Brown, N. Carlini, C. Zhang, C. Olsson, P. Christiano, and I. Goodfellow, "Unrestricted adversarial examples," *arXiv preprint arXiv:1809.08352*, 2018.

[55] Y. Song, R. Shu, N. Kushman, and S. Ermon, "Constructing unrestricted adversarial examples with generative models," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[56] N. Inkawhich, M. Inkawhich, Y. Chen, and H. Li, "Adversarial attacks for optical flow-based action recognition classifiers," *arXiv preprint arXiv:1811.11875*, 2018.

[57] R. Pony, I. Naeh, and S. Mannor, "Over-the-air adversarial flickering attacks against video recognition networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[58] Z. Chen, L. Xie, S. Pang, Y. He, and Q. Tian, "Appending adversarial frames for universal video attack," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 3199–3208.

[59] D. Kumar, C. Kumar, C. W. Seah, S. Xia, and S. Ming, "Finding achilles' heel: Adversarial attack on multi-modal action recognition," in *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 2020.

[60] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 2001, pp. 327–340.

[61] L. A. Gatys, A. S. Ecker, and B. Matthias, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576, 2015*, 2015.

[62] J. Justin, A. Alexandre, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[63] H. Huang, H. Wang, W. Luo, L. Ma, W. Jiang, X. Zhu, Z. Li, and W. Liu, "Real-time neural style transfer for videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 783–791.

[64] C. Gao, D. Gu, F. Zhang, and Y. Yu, "Reconet: Real-time coherent video style transfer network," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2018, pp. 637–653.

[65] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," in *Proceedings of the German Conference on Pattern Recognition (GCPR)*, 2016, pp. 26–36.

[66] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[67] P. Heckbert, "Color image quantization for frame buffer display," *ACM Siggraph Computer Graphics*, vol. 16, no. 3, pp. 297–307, 1982.

[68] A. R. Smith, "Color gamut transform pairs," *ACM Siggraph Computer Graphics*, vol. 12, no. 3, pp. 12–19, 1978.

[69] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[70] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 2230–2236.

[71] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, pp. 1135–1151, 1981.

[72] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[73] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[74] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2011, pp. 2556–2563.

[75] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, 2004.

[76] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of Multivariate Analysis*, 1982.

[77] "Amazon mechanical turk," https://www.mturk.com.

[78] H. Mann and W. D.R., "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, 1947.

[79] G.-H. Lee, Y. Yuan, S. Chang, and T. Jaakkola, "Tight certificates of adversarial robustness for randomly smoothed classifiers," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[80] G. Yang, T. Duan, J. E. Hu, H. Salman, I. Razenshteyn, and J. Li, "Randomized smoothing of all shapes and sizes," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119, 2020, pp. 10 693–10 705.

[81] A. Kumar, A. Levine, T. Goldstein, and S. Feizi, "Curse of dimensionality on randomized smoothing for certifiable robustness," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[82] A. Blum, T. Dick, N. Manoj, and H. Zhang, "Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images," *Journal of Machine Learning Research (JMLR)*, 2020.

[83] Y. Bar, N. Levy, and L. Wolf, "Classification of artistic styles using binarized features derived from a deep neural network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[84] T. Sun, Y. Wang, J. Yang, and X. Hu, "Convolution neural networks with two pathways for image style recognition," *IEEE Transactions on Image Processing (TIP)*, 2017.

[85] E. Cetinic, T. Lipic, and S. Grgic, "Fine-tuning convolutional neural networks for fine art classification," *Expert Systems with Applications (ESWA)*, 2018.

[86] C. B. El Vaigh, N. Garcia, B. Renoust, C. Chu, Y. Nakashima, and H. Nagahara, "Gcnboost: Artwork classification by label propagation through a knowledge graph," in *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*, 2021.

[87] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2013, pp. 1385–1392.

[88] R. Likert, "A technique for the measurement of attitudes." *Archives of Psychology*, 1932.
[89] K. Hara, A. Adams, K. Milland, S. Savage, C. Callison-Burch, and J. P. Bigham, "A data-driven analysis of workers' earnings on amazon mechanical turk," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI)*, 2018, pp. 1–14.

APPENDIX

A. STYLEFOOL CONFIGURATION

Following Huang *et al.* [63], the output of layer `Relu4_2` is used in content loss, and the outputs of layers {`Relu1_1`, `Relu2_1`, `Relu3_1`, `Relu4_1`, `Relu5_1`} are used for style loss. The selected layers are all replaceable, depending on personal preference. We use DeepFlow [87] in optical flow estimation.

In style selection, we set bottom circle radius $r = 50$, height $h = 50\sqrt{3}$, weight coefficient $\mu = 10^4$. In style transfer, we set weight coefficients $\alpha = 10$, $\beta = 75$ for targeted attacks and $\beta = 50$ for untargeted attacks, $\gamma = 10^{-3}$, $\lambda = 10^3$. The weight coefficients μ, α, β, γ, λ are fine-tuned on 25 videos that are randomly selected from the UCF-101 dataset under C3D model. A grid search is carried on to find the most reasonable values. The current parameter combination is a reference with excellent performance. Users can slightly modify the parameters as needed. In gradient estimation, we set the number of Gaussian noise $n = 64$, the noise variance $\sigma = 10^{-6}$ for targeted attacks, 10^{-3} for untargeted attacks.

B. DETAILS OF USER STUDY

Survey protocol. Considering that the naturalness test and the realness test evaluate the indistinguishability of videos from different angles, it could be difficult for participants to make a fair enough decision in a short time period. In order to avoid bias, we evenly separate participants into 2 groups (80 participants each for the naturalness test and the realness test). In the naturalness test, 20 randomly selected videos, including 10 adversarial samples and 10 clean videos, are presented to each participant. The participants are required to evaluate the naturalness after watching each video, and then judge the sensory comfort according to their cognition and common sense, based on a Likert scale [88] from 1 to 5, representing "very unnatural", "somehow unnatural", "neutral", "somehow natural", and "very natural", respectively. In the realness test, 10 pairs of adversarial videos and their corresponding clean videos are displayed simultaneously to each participant who is asked to identify the realness of two videos (*i.e.,* which video is a clean video). Note that, participants are not informed of how many videos in each video pair are real, *i.e.,* we designed a multiple choice question with 5 answers, "Video 1 real and Video 2 unreal", "Video 1 unreal and Video 2 real", "both real", "both unreal", and "unable to judge". To filter out low quality or random responses, the answers that are made in less than 10 seconds (inclusive of the video playing time) are ignored. The average video duration is 7 seconds, and we have reserved at least 3 seconds for thinking.

Detailed procedure. We conducted the user study on the naturalness and realness aspects of the adversarial videos of StyleFool with 160 native speaker participants in total recruited on Amazon Mechanical Turk, a crowdsouring platform to hire remotely located "crowdworkers" to perform discrete on-demand tasks that computers are currently unable to do. It is operated under Amazon Web Services, and is owned by Amazon. In our user study, we only hired participants over 18 years old who speak English as the first language and are from USA, UK and Australia with an answer approval rate of at least 95%. A computer related background is not necessary for participants. We paid each participant an average of $20.00/hr (*i.e.,* $0.9 per question), which is higher than the average payment ($11.00/hr) on the platform [89].

C. ADDITIONAL EXPERIMENTAL RESULTS

Figure 5 shows the original video samples and the corresponding adversarial video samples generated by StyleFool. Examples include both targeted and untargeted attacks. The styles of video samples are changed, but the semantic information remains unchanged. Please see more details in Section IV-B.

We report the quantitative results of the indistinguishability in Table IX, including the SSIM, PSNR, and FID between original videos and adversarial videos (Ilyas [35], V-BAD [20], H-Opt [21] and "StyleFool-ori-adv"), or between stylized videos and adversarial videos ("StyleFool-sty-adv"). Please see more details in Section IV-B.

Figure 6 shows the original video samples and the corresponding adversarial video samples generated by StyleFool under different ε_{adv}. Examples include both targeted and untargeted attacks. Please see more details in Section IV-D.

Figure 7 shows the original video samples and the corresponding adversarial video samples generated by StyleFool under different style selection methods in the ablation study. Examples include both targeted and untargeted attacks. Please see more details in Section IV-D.

Figures 8 and 9 provide more visualization results. In Figure 8, different target labels are selected in targeted attacks. The original video clip is shown in the first column. The style images, shown in the first row, are selected according to the criterion in Equation 5. The corresponding adversarial video clips are shown below the style images.

In Figure 9, different styles are selected in untargeted attacks. Images in the first row represent the selected style images, while images in the second row stand for the original video clip and adversarial video clips. The number below the label name represents the ranking of the criterion when selecting the style. Please note that the ranking is not a continuous number because similar style images are hidden in order to visualize more adversarial examples. Note that, for untargeted attacks, only the sample with the smallest criterion has the best effect.

Table X shows the time cost in style set preparation. We also provide the attack time costs of StyleFool and benchmark attack methods with the same set of 100 randomly-selected videos in Table XI. However, the number of queries is the main evaluation criterion for black-box adversarial attacks including

(a) Targeted Attacks

(b) Untargeted Attacks

Fig. 5: Visualization of StyleFool. The left four screenshots are taken from original videos, the right four screenshots are taken from adversarial videos, the image in the middle is the selected style image.

TABLE IX: Quantitative analysis on video indistinguishability.

Model	Attack	UCF-101 (Targeted)			HMDB-51 (Targeted)			UCF-101 (Untargeted)			HMDB-51 (Untargeted)		
		SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓
C3D	Ilyas [35]	0.7610	27.6885	136.4496	0.7945	27.5548	122.5155	0.7895	30.7324	112.7122	0.7910	30.5597	107.1496
	V-BAD [20]	0.7863	28.4952	119.1019	0.7948	29.3521	113.6902	0.7783	30.5321	108.0545	0.7873	30.5421	87.4092
	H-Opt [21]	0.6768	23.4829	117.7430	0.7388	26.0329	109.1304	0.7144	30.6636	104.3397	0.7972	37.4704	102.6211
	StyleFool-ori-adv	0.4142	12.9681	215.6654	0.4149	12.8772	205.9897	0.5174	16.4794	156.4410	0.5786	18.8988	128.8778
	StyleFool-sty-adv	**0.8653**	**29.4112**	**73.2175**	**0.8496**	**30.9609**	**97.9852**	**0.8735**	**45.2092**	**53.7769**	**0.9262**	**64.9507**	**30.7555**
I3D	Ilyas [35]	0.6690	29.6478	136.0133	0.7100	29.5331	125.4818	0.6851	30.6850	124.6638	0.7075	30.5364	110.9063
	V-BAD [20]	0.7199	29.2051	148.0453	0.7252	29.0609	136.1517	0.6861	30.5674	149.1361	0.6941	30.5106	132.6454
	H-Opt [21]	0.7044	29.7635	163.5481	0.7255	28.5815	128.9625	0.7262	29.3567	141.4572	0.7742	31.6156	126.8154
	StyleFool-ori-adv	0.4009	13.4977	216.6966	0.4060	14.9125	222.1958	0.5113	16.3142	194.2943	0.4865	17.6216	187.8469
	StyleFool-sty-adv	**0.7773**	**31.2119**	**104.2144**	**0.7607**	**30.3052**	**102.2426**	**0.8464**	**42.4192**	**102.1841**	**0.8092**	**46.6260**	**88.9165**

(a) Targeted Attack

(b) Untargeted Attack

Fig. 6: Visualization of StyleFool results with different ε_{adv}: (a) Targeted attack ("Kick" to "Talk"); (b) Untargeted attack ("BlowingCandles" to "PlayingFlute").

StyleFool. The time cost is not a major concern as long as the model can be attacked through fewer queries. Please note that StyleFool requires extra time (about 21-28 mins) for style transfer.

TABLE X: Preparation time cost per image.

Generating styles	Calculating color proximity	Calculating confidence score	Total
0.0005s	0.3864s	0.0305s	0.4174s

TABLE XI: Attack time cost.

Model	Attack	Time costs (min)			
		UCF101 (Targeted)	HMDB51 (Targeted)	UCF101 (Untargeted)	HMDB51 (Untargeted)
C3D	Ilyas [35]	105	74	17	7
	V-BAD [20]	38	59	6	15
	H-Opt [21]	188	224	20	53
	StyleFool	10	10	2	<1
I3D	Ilyas [35]	66	58	18	6
	V-BAD [20]	32	28	16	5
	H-Opt [21]	46	53	28	20
	StyleFool	10	11	1	<1

Style transfer has demonstrated substantial efficacy in the search of the potential decision boundary of the target model. We compare the boundary search using style transfer and that with a surrogate model. Since we are the first one to propose unrestricted perturbations in videos, we can only compare style transfer with models with restricted perturbations. We randomly choose 100 videos from the dataset, and attack them using V-BAD against I3D model. These adversarial

(a) Original (e) Both - targeted StyleFool

(b) Randomly (f) Original

(c) Color theme proximity (g) Randomly

(d) Confidence score (h) Color theme proximity - untargeted StyleFool

Fig. 7: Visualization of StyleFool results with different style selection strategies. (a)-(e) Targeted attacks; (f)-(h) Untargeted attacks.

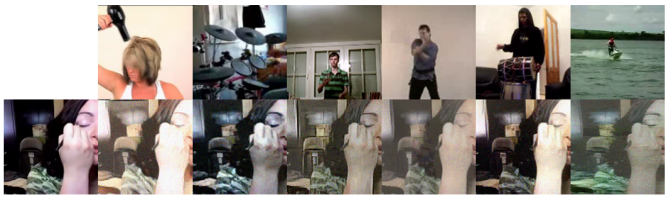


Fig. 8: Adversarial examples with different target labels for targeted attacks.

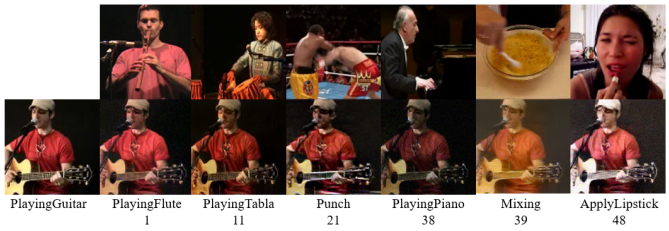


Fig. 9: Adversarial examples with different styles for untargeted attacks.

TABLE XII: Attack performance of videos initiated by different methods.

Model	Attack	UCF-101 (Targeted)				HMDB-51 (Targeted)			
		ASR	minQ	maxQ	AQ	ASR	minQ	maxQ	AQ
C3D	V-BAD+I3D	100	2,538	288,454	130,116	100	205	272,999	75,756
	StyleFool	**100**	**1,284**	**248,131**	**65,066**	**100**	**101**	**95,897**	**35,232**
Model	**Attack**	**UCF-101 (Untargeted)**				**HMDB-51 (Untargeted)**			
		ASR	**minQ**	**maxQ**	**AQ**	**ASR**	**minQ**	**maxQ**	**AQ**
C3D	V-BAD+I3D	100	1	79,969	10,020	100	1	30,148	3,732
	StyleFool	**100**	**1**	**19,013**	**3,811**	**100**	**1**	**9,948**	**1,526**

videos are again attacked using the same adversarial samples generation method of StyleFool against C3D model. As shown in Table XII, the results show that StyleFool can reduce at least half of the queries in both targeted and untargeted attacks, which means style transfer performs better in pushing videos to the decision boundary. The results also indicate that video adversarial samples lack transferability since large number of queries are still needed for attacks.

D. HSV-to-XYZ Transformation

In the HSV space, *i.e.,* the HSV cone, let r be the bottom radius and h be the height of the cone. To compute the similarity between two pixels in the HSV space, we first convert an HSV coordinate (H, S, V) to a coordinate (X, Y, Z) in the XYZ space as follows:

$$\begin{cases} X = rVS\cos H \\ Y = rVS\sin H \\ Z = h\,(1-V)\,. \end{cases} \tag{13}$$

Given pixel values of the i-th color theme (H_i^x, S_i^x, V_i^x) of the clean video x and the j-th color theme (H_j^s, S_j^s, V_j^s) of a style image s, we obtain their coordinates $\phi_i^x = (X_i^x, Y_i^x, Z_i^x)$ and $\phi_j^s = (X_j^s, Y_j^s, Z_j^s)$ in the XYZ space, respectively.

TABLE XIII: SSIM analysis between frames.

Frame index	1st and 2nd	1st and the middle	1st and the last
Average SSIM	0.5948	0.5411	0.5100

E. SSIM Analysis between Different Frames

As discussed in Section VI, we currently focus on 16-frame videos. Table XIII shows an SSIM analysis between different frames, which indicates that the difference between the first frame and other frames in a single-action video is limited.