applied sciences

MDPI

Article

A Robust Adversarial Example Attack Based on Video Augmentation

Mingyong Yin, Yixiao Xu, Teng Hu and Xiaolei Liu *

Institute of Computer Application, China Academy of Engineering Physics, Mianyang 621900, China
* Correspondence: caeplxl@126.com

Abstract: Despite the success of learning-based systems, recent studies have highlighted video adversarial examples as a ubiquitous threat to state-of-the-art video classification systems. Video adversarial attacks add subtle noise to the original example, resulting in a false classification result. Thorough studies on how to generate video adversarial examples are essential to prevent potential attacks. Despite much research on this, existing research works on the robustness of video adversarial examples are still limited. To generate highly robust video adversarial examples, we propose a video-augmentation-based adversarial attack (v3a), focusing on the video transformations to reinforce the attack. Further, we investigate different transformations as parts of the loss function to make the video adversarial examples more robust. The experiment results show that our proposed method outperforms other adversarial attacks in terms of robustness. We hope that our study encourages a deeper understanding of adversarial robustness in video classification systems with video augmentation.

Keywords: video augmentation; adversarial example; video classification

Citation: Yin, M.; Xu, Y.; Hu, T.; Liu X. A Robust Adversarial Example Attack Based on Video Augmentation. *Appl. Sci.* **2023**, *13*, 1914. https://doi.org/10.3390/app13031914

Academic Editor: Luis Javier García Villalba

Received: 20 October 2022
Revised: 6 January 2023
Accepted: 27 January 2023
Published: 1 February 2023

1. Introduction

With the development of deep neural network technologies, deep-learning-based models and algorithms have been applied to various scenarios, including some security- and safety-critical conditions (e.g., autonomous driving, face authentication, and malicious application detection). However, misclassification caused by the crashes of deep-learning models may lead to systematic security risks such as traffic incidents and crimes against deep-learning-based economic systems. Therefore, evaluating and improving the robustness of deep learning models against potential security threats has become a focus of recent studies.

Over the past few years, a number of studies have shown evidence that existing deep learning systems are susceptible to attacks using adversarial examples [1–4]. The malicious users of smart systems construct minor, human-indiscernible perturbations and apply them to benign inputs such as photographs and audio data. This leads to model misclassifications, and the inputs that have been manipulated in this manner are referred to as adversarial instances. As a result of the opaque nature of deep learning models and their limited capacity for interpretation, adversarial assaults have emerged as a fundamental challenge to their development.

To overcome the threat raised by adversarial examples, adversarial training is now recognized as one of the most efficient defense methods against adversarial attack [5–7]. However, the effectiveness of adversarial training is highly related to the similarity of training adversarial examples and real-world adversarial examples. How to generate robust adversarial examples has a significant impact on the adversarial training effect. Currently, the majority of the research on robust adversarial examples are in the image classification field [8–11] and the speech recognition field [12–16], leaving the video classification field unexplored.

Generating video adversarial examples (Figure 1) is another vital field where robustness is essential to the defense training methods [17]. However, it is much more difficult to generate

robust adversarial examples for video than for image or audio. Compared with image and audio data, video data consists of temporal and spatial information at the same time. To generate robust video adversarial examples, there are still several challenges to be addressed:

(C1) A common method used to defend against video adversarial attacks is to add small noises or distortions to the input examples. This is done because even minute differences have the potential to corrupt the semantic information carried by adversarial perturbations. Since adversarial perturbations are typically very small and difficult for humans to detect, this defense method is effective. However, video adversarial examples that are built using traditional methods have a very low level of robustness and, for the most part, lose their adversarial properties when subjected to these defense methods.

(C2) Adding noise in the construction process can escape the defense (mentioned in C1) and increase the robustness of adversarial examples to some extent. However, it will increase the perturbation of video samples significantly, which weakens the human imperceptibility of video adversarial examples.

(C3) The defense method of video adversarial examples can be improved by simply changing the video processing strategy, thus bypassing the robustness enhancement method (mentioned in C2).

To overcome the mentioned difficulties, we propose a **v**ideo-**a**ugmentation-based **a**dversarial **a**ttack (**v3a**), which is able to generate robust video adversarial examples in white-box or black-box scenes.

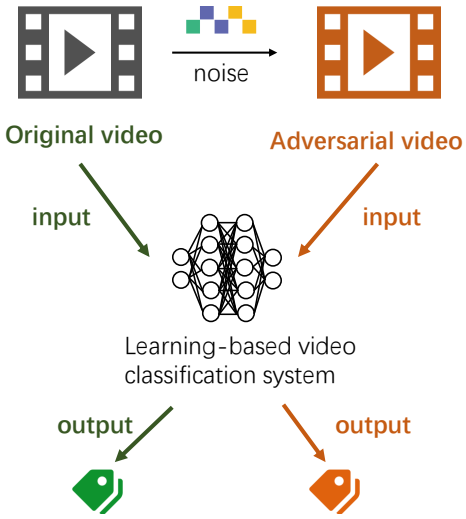

Figure 1. General process of video adversarial example attack.

Combined with the video augmentation processing, v3a improved the loss function and thus generates more robust video adversarial examples which can escape the adversarial defense (addressing C1).

Meanwhile, by performing once-transformation before generating and only transforming the adversarial perturbations in each iteration, v3a can increase the robustness of video adversarial examples without noticeable perturbations (addressing C2).

In addition, we investigate the influence of different transformations as parts of the loss function to generate video adversarial examples (addressing C3).

Finally, experiments on the UCF-101 dataset demonstrate the robustness of the adversarial examples generated by our method, which substantially improves the robustness of existing methods.

2. Materials

Video adversarial attacks. Video is composed by superimposing multiple frames of images, so initially, some scholars directly migrated the approach of image adversarial attack to video [18–20]. However, considering the video has information in both time and

space domains, it is necessary to construct video adversarial examples in a larger search space, which makes video adversarial attacks more challenging. By selecting keyframes of the video and perturbing them, Wei et al. [21] proposed the first sparse attack method (SA) to improve the efficiency of video adversarial attacks. However, in SA, the keyframe selection process is independent of the victim model and the input example, therefore, the effect of the selected keyframes in SA is limited. To address this issue, Wei et al. [22] proposed a difference-based heuristic approach to achieve effective keyframe selection. The heuristic method measures the importance of a frame by comparing the difference in output when inputting two sets of frames containing the particular frame or not. Later, Xu et al. [23] used Grad-CAM to further improve the accuracy and efficiency of keyframe selection. In addition to studying how to improve construction efficiency, some researchers [24–26] also focus on how to perform black-box video adversarial attacks to improve their operability. Patch Attack (V-BAD) is the first black-box video adversarial attack method to simplify the black-box gradient estimation process [24]. Patch Attack attempts to reduce the dimensions of search space for gradient estimation by treating each frame as a separate image and estimating the overall gradient of the pixel points in a region. Later, Motion-Sampler Attack further reduced the dimensions of search space by obtaining a motion-aware perturbation prior to guiding the black-box gradient estimation, resulting in fewer queries [27]. However, the fixed temporal structure of the search space is not flexible enough. Thus, Geo-Trap Attack, instead of fixing the temporal structure of the search space, parameterizes the temporal structure of gradients with geometric transformations [26].

However, to the best of our knowledge, none of these above-mentioned video adversarial attack methods considers the robustness of the generated adversarial examples. Our proposed method can be applied to all these current methods to achieve a trade-off between efficiency, operability, and robustness.

Robustness of adversarial examples. Despite the initial success of adversarial attacks toward different deep learning models. Existing adversarial attack methods that mainly focus on the attack success rate are usually not robust due to the inevitable noise and biases in the transmission process [28]. Thus, some recent studies explored the robustness enhancement methods for adversarial attacks. Luo et al. [28] introduced the human perceptual system as well as the noise tolerance evaluation principle into the image adversarial attack process. By iteratively enhancing the noise tolerance of adversarial examples, the robustness of the proposed method is significantly improved. Athalye et al. [29] further considered the robustness of image adversarial examples under different transformations and achieved physical attacks by generating transformation-robust adversarial examples. Eykholt et al. [30] introduced more types of transformations that simulate the effects of the physical world on image adversarial examples and achieved physical attacks with a higher success rate and smaller perturbation size. However, these adversarial attack robustness enhancement methods are designed for image adversarial examples. Thus, existing methods usually suffer from a large perturbation size and poor imperceptibility given the property that video examples have higher dimensionality compared with image examples.

In this work, we introduce the augmentation-based robustness enhancement method to video adversarial attacks, and propose v3a, a video-augmentation-based adversarial attack. Moreover, we significantly reduce the extra perturbations caused by iterate transformations.

3. Methodology

To overcome the limitations in the robustness and to address the above-mentioned challenges of existing available video adversarial attack methods, in this section, we propose the video-augmentation-based adversarial attack method, v3a.

3.1. Problem Formulation

Given the target video classification model F, the input examples can be represented as $\boldsymbol{x} \in \mathbb{R}^{N \times C \times H \times W}$, where $NCHW$ denote the index of frames, the number of color channels, and the height and the width of each frame, respectively. $y = F(\boldsymbol{x})$ represents the inference

result of the video classification model F. When performing untargeted adversarial attacks, the goal of the adversary is to generate an adversarial perturbation δ, which misleads the victim model F into a wrong prediction ($y \neq F(\boldsymbol{x} + \boldsymbol{\delta})$). For targeted attacks, the goal of the adversary is to generate an adversarial perturbation δ, which forces F to make the prediction of the specified target label $\hat{y}$ ($\hat{y} = F(\boldsymbol{x} + \boldsymbol{\delta})$). The robustness and the imperceptibility of adversarial examples usually show a trade-off trend because larger perturbation sizes often lead to better robustness but poorer imperceptibility. Generally, targeted attacks require more attack iterations and larger perturbation sizes, thus targeted adversarial examples are more robust.

Regarding the restrictions placed on the attack, we will assume that it is carried out under both white-box and black-box conditions. In the white-box scenario, the adversary is in possession of all of the information regarding the target model, which includes the structure and parameters of the model, and is, therefore, able to directly compute the gradient. On the other hand, when the condition is a black box, the adversary is only able to estimate the gradient by referring to the output of the model. Black-box video adversarial examples are said to have larger perturbation sizes compared to adversarial examples that were generated under white-box conditions, as stated by existing video adversarial attack methods [18–20].

3.2. Adversarial Video Attack

In this paper, we use the sparse adversarial video attack method (SA) [21] and the heuristic black-box adversarial video attack method (HA) [22] as two baseline methods of our proposed method, v3a. SA is the first sparse adversarial attack method for video classification models which achieved key-frame selection under white-box constraints. By selecting a constant number of frames and only adding adversarial perturbations on these frames, SA accelerates video adversarial attacks and reduces the perturbation sizes. HA first extended SA to black-box conditions with a zero-order gradient estimation algorithm. To perform key-frame selections under black-box conditions, HA further proposes a heuristic method that helps evaluate the importance of different frames. Following SA and HA, the optimization process of untargeted attacks of v3a can be described as:

$$\underset{\delta}{argmin}\ \alpha||\boldsymbol{M} \cdot \boldsymbol{\delta}||_{2,1} - \beta L(y, F_\theta(\boldsymbol{x} + \boldsymbol{M} \cdot \boldsymbol{\delta})) \tag{1}$$

where $\boldsymbol{M} \in \{0,1\}^{N \times C \times H \times W}$ denotes the sparse mask which achieves sparse attack via keyframe selection. For a specific frame in the input example $\boldsymbol{x}$, the value of the corresponding position of $\boldsymbol{M}$ is set to 1 and the other positions are set to 0. $L(.,.)$ is the adversarial loss function that calculates the cross-entropy loss. F_θ denotes the victim model with parameters θ. $||\boldsymbol{\delta}||_{2,1}$ calculates the $L_{2,1}$ norm for the generated perturbation which serves to reduce the perturbation size. (α, β) are the two weights of the different terms of the equation.

As for targeted attacks, the optimization process can be represented as follow:

$$\underset{\delta}{argmin}\ \alpha||\boldsymbol{M} \cdot \boldsymbol{\delta}||_{2,1} + \beta L(\hat{y}, F_\theta(\boldsymbol{x} + \boldsymbol{M} \cdot \boldsymbol{\delta})) \tag{2}$$

where $\hat{y}$ denotes the target label of the attack.

For white-box conditions, attackers use gradient-based optimization methods to solve the above-defined optimization problems by directly computing the gradient. However, the attacker cannot directly compute the gradients under black-box conditions, so HA uses the zero-order algorithm for black-box gradient estimation. The estimated gradient is defined as:

$$\hat{\boldsymbol{g}} = \frac{1}{N} \sum_{n=1}^{N} \frac{F(\boldsymbol{x} + \boldsymbol{\delta} + \mu \boldsymbol{\phi_n}) - F(\boldsymbol{x} + \boldsymbol{\delta})}{\mu} \boldsymbol{\phi_n} \tag{3}$$

where $\boldsymbol{\phi_n}$ is sampled from the standard Gaussian distribution and N is the number of samples. μ is a smoothing parameter, we set μ to 0.005 in the experiment. During each iteration, the update of δ is given by

$$\delta \leftarrow \delta - s\hat{g} \tag{4}$$

where s denotes the step size.

Meanwhile, HA assumes that the victim black-box model only provides the top-1 prediction label, thus there is no available information for the adversary to guide the gradient descent process. So HA introduces the boundary attack algorithm [31], which initializes the adversarial perturbation by randomly selecting a sample from the target class and then iteratively reduces the perturbation size via gradient descent algorithms. Therefore, the optimization goal is changed to minimize the perturbation size while keeping adversarial properties. The optimization process of boundary attack can be formulated as:

$$d(\boldsymbol{x}, \hat{\boldsymbol{x}}^{\boldsymbol{k}}) = d(\boldsymbol{x}, \hat{\boldsymbol{x}}^{\boldsymbol{k-1}} + \eta^k) + \epsilon \cdot d(\boldsymbol{x}, \hat{\boldsymbol{x}}^{\boldsymbol{k-1}}) \tag{5}$$

where d denotes the distance of the original video x and the corresponding adversarial example $\hat{x}$, k denotes the k-th iteration, η denotes the random direction extracted by a binary search process, and ϵ denotes the smoothing parameter.

In this paper, we use SA and HA as two baseline methods that mainly focus on the attack success rate but rarely consider the robustness of the generated adversarial examples. Meanwhile, SA and HA can represent black-box attack methods. Additionally, our proposed robustness enhancement method v3a is based on SA and HA. v3a can also be applied to other existing video adversarial attack methods easily and enhances the robustness of adversarial examples.

3.3. Video Augmentations Based Adversarial Attack

Adversarial examples may introduce noise or image transformations during delivery, thus the robustness of the examples has a large influence on the success rate. At the same time, smart system owners may introduce noise or transformations for security considerations. Existing adversarial video attacks did not consider the potential noise and variation, resulting in a low success rate and poor robustness. To address this challenge, We propose the video-augmentation-based adversarial attack method v3a. By introducing noise and video-augmentation-based transformations during the attack process, our method significantly improves the robustness of adversarial examples while keeping an acceptable perturbation size. Figure 2 gives an overview of v3a. Instead of applying adversarial perturbations to the clean input video frames directly, v3a performs video-augmentation-based transformations on the generated perturbations and the original frames.

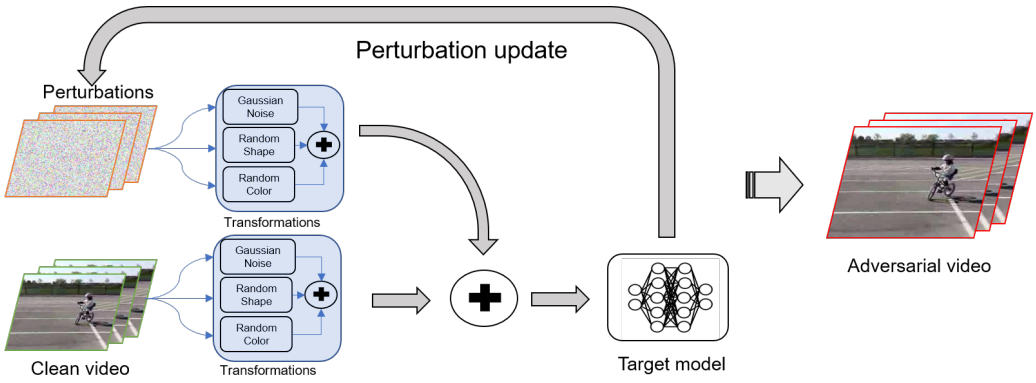

Figure 2. Overview of the video-augmentation-based adversarial attack.

Specifically, we consider three factors that may affect the adversarial example: Gaussian noise, video shape transformation, and video color transformation. Gaussian noise

may be the most common transformation that may occur during the transmission process of video data and is also easy to be introduced by security defenders. Video shape transformations include random rotation, expansion, and contraction, and video color transformations include random brightness and saturation adjustment. These factors are also likely to occur during video transmission. They have little effect on the classification of clean examples because they have little impact on the semantic information of the clean examples. However, these transformations will lead to the failure of adversarial perturbations with high probabilities. Figure 3 gives a visualization of three different types of video augmentations introduced in our method. Given the input $\boldsymbol{x}$ and three transformations denoted by T_1, T_2, T_3, respectively, the transformation function of our method can be defined as follows:

$$T(\boldsymbol{x}) = \gamma_1 T_1(\boldsymbol{x}) + \gamma_2 T_2(\boldsymbol{x}) + \gamma_3 T_3(\boldsymbol{x}) \tag{6}$$

where T represents the transformation function, $\gamma_1, \gamma_2, \gamma_3 > 0$ and $\gamma_1 + \gamma_2 + \gamma_3 = 1$, balancing three different transformations.

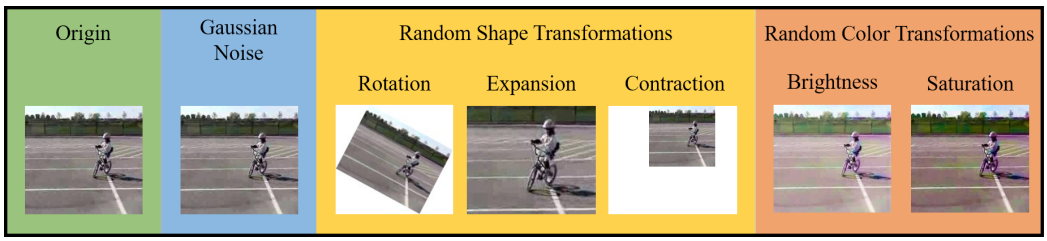

Figure 3. Visualization of different video augmentations introduced in v3a.

Meanwhile, we consider three different ways of adding transformations to the iterations:

- **Before iterations.** By performing once-transformation before the attack, the robustness of adversarial examples can be improved to some degree. This method is fast but limited because it does not consider the effect of the perturbation.
- **During iterations.** By performing transformation in each iteration, the robustness of adversarial examples can be greatly enhanced. However, repeated transformations can make attacks more difficult or even fail, as the original examples transform multiple times. At the same time, the size of the perturbation will keep increasing with iterations, which leads to weaker concealment.
- **During iterations but only for perturbations.** By performing once-transformation before the attack and transforming the adversarial perturbations in each iteration, the robustness of adversarial examples is effectively improved and the difficulty of the attacks does not increase significantly. This approach thus meets the comprehensive needs.

Therefore, the optimization goal of v3a can be described as follows:
For untargeted attacks,

$$\underset{\delta}{argmin}\ \alpha||\boldsymbol{M} \cdot T(\delta)||_{2,1} - \beta L(y, F_\theta(T(\boldsymbol{x} + \boldsymbol{M} \cdot \delta))) \tag{7}$$

where the start point of the algorithm is changed from the benign input video $\boldsymbol{x}$ to the once-transformed video example $T(\boldsymbol{x})$.

For targeted attacks,

$$\underset{\delta}{argmin}\ \alpha||\boldsymbol{M} \cdot T(\delta)||_{2,1} + \beta L(\hat{y}, F_\theta(T(\boldsymbol{x} + \boldsymbol{M} \cdot \delta))) \tag{8}$$

Meanwhile, our black-box gradient estimation is given by the following equation:

$$\hat{\boldsymbol{g}} = \frac{1}{N}\sum_{n=1}^{N} \frac{F(T(T(\boldsymbol{x}) + \delta + \mu\boldsymbol{\phi_n})) - F(T(T(\boldsymbol{x}) + \delta))}{\mu}\boldsymbol{\phi_n} \tag{9}$$

Algorithm 1 show the implementation of v3a.

Algorithm 1: v3a: video-augmentation-based adversarial attack.

Input: target model F;
clean video x, clean label y;
transformation function T;
search distribution $\pi(\theta|z)$;
target class video x^*, target label $\hat{y}$ //if targeted attack
Output: adversarial example $\hat{x}$

if *black-box* **then**
  $\delta \leftarrow x^* - x$;
end
for *i in max_iter* **do**
  if *white-box* **then**
    $g \leftarrow \nabla_{\delta_{white}} F(T(x) + T(\delta))$;
  end
  if *black-box* **then**
    $g^* \leftarrow \nabla_{\delta_{black}} D(T(x) + T(\delta))$;
  end
  if *white-box* **then**
    $\delta \leftarrow \delta + \eta g$;
  end
  if *black-box* **then**
    $\delta \leftarrow B(T(\delta - \eta \hat{g^*}))$;
  end
end
$\hat{x} \leftarrow T(x + \delta)$;
return $\hat{x}$

4. Experimental Results

4.1. Experimental Methodology

Dataset. We evaluate the performance of our method v3a on the most widely used video classification dataset, UCF-101 [32]. UCF-101 is a dataset containing 13,320 videos including 101 action classes for action recognition. During the evaluation, we randomly select one video from each class and sample 40 frames for each video as the input stream of the target video classification model.

Target model. We use a long-term recurrent convolutional network (LRCN) [33] as the victim model, for instance, SA [21]. The LRCN model consists of two parts: (1) a CNN part for feature maps extracting from the original frames, and (2) a recurrent neural network (RNN) part for prediction making.

Metrics. We evaluate our approach with two metrics: (1) Fooling rate (F). The percentage of success adversarial examples [34]. In this paper, the fooling rate is evaluated under video transformations. We use the fooling rate to evaluate the ability of v3a that can bypass video-augmentation-based transformations defense. (2) Mean absolute perturbation (MAP). The mean absolute perturbation size of each pixel throughout the video. We use the mean absolute perturbation to evaluate the impact of v3a on the perturbation sizes, which also represents the human imperceptibility of different attack methods.

Attack settings. We consider both untargeted attacks and targeted attacks under white-box conditions and black-box conditions. Note that we have three parameter weight $\gamma_1\gamma_2\gamma_3$ in transformation function. The parameters balance the effect of each transformation on the perturbations. Thus, we determined the appropriate values for $\gamma_1\gamma_2\gamma_3$ through a grid search as $(0.5, 0.25, 0.25)$, respectively.

4.2. Performance Comparison

We compare our method to two other adversarial video attack methods: the sparse adversarial video attack (SA) [21], and the heuristic black-box adversarial video attack (HA) [22] for white-box conditions and black-box conditions, respectively. Both of these methods are used to attack videos in a variety of environments. We examined the effectiveness of three different approaches to attack while they were undergoing transformations.

4.2.1. Performance Comparison of White-Box Attacks

Table 1 shows the performance comparison of our method and SA under white-box conditions. According to the results, when no transformation is performed on the adversarial examples, both SA and v3a achieved a 100% success rate, which demonstrates that v3a will not weaken the adversarial properties of adversarial examples under benign conditions. When performing transformations on video adversarial examples, v3a achieves a higher fool rate than SA, demonstrating that our method can effectively improve the robustness of adversarial examples against potential transformations. For both targeted and untargeted attacks, the fool rates of our method are 12% to 27% higher than SA against three single transformations. By comparing the difference in success rates, it can be found that Gaussian noise has the greatest effect on adversarial examples, and the ensemble of transformations is more powerful compared to the individual transformations. Adversarial examples of targeted attacks are more susceptible to transformations than those of untargeted attacks. At the same time, the comparison of the MAP of the two methods illustrates that our method does not significantly magnify the perturbations. By only transforming the perturbations, our method has an acceptable impact on the quality of adversarial examples.

4.2.2. Performance Comparison of Black-Box Attacks

For black-box attacks, the experimental results showed a similar tendency. When no transformation is performed on the adversarial examples, both HA and v3a achieved a 100% success rate. When performing transformations on video adversarial examples, v3a achieves a higher fool rate than HA. As shown in Table 2, by introducing the transformations into the black-box gradient estimation process, our method improves the robustness of black-box adversarial examples. On the UCF-101 dataset, our method improves the success rate by more than 5% for all three individual and ensemble transformations. Meanwhile, it can be observed that black-box adversarial examples are more robust than white-box ones because the MAP is larger.

Table 1. White-box attacks against the LRCN model. The MAP of SA and our method are 0.0125 and 0.0133, respectively.

Transformations	Method	FR Untargeted	FR Targeted
Benign	SA	100%	100%
	v3a	100%	100%
Gaussian Noise	SA	72.28%	65.35%
	v3a	97.03%	93.07%
Random Shape	SA	81.19%	76.24%
	v3a	95.05%	91.09%
Random Color	SA	85.70%	80.20%
	v3a	98.02%	96.40%
Assemble	SA	57.14%	49.50%
	v3a	89.11%	84.16%

Table 2. Black-box attacks against the LRCN model. The MAP of HA and our method are 1.4528 and 1.4991, respectively.

Transformations	Method	FR Untargeted	FR Targeted
Benign	HA	100%	100%
	v3a	100%	100%
Gaussian Noise	HA	89.11%	84.16%
	v3a	94.06%	90.10%
Random Shape	HA	91.09%	88.12%
	v3a	96.04%	94.06%
Random Color	HA	87.13%	87.13%
	v3a	98.02%	93.07%
Assemble	HA	80.20%	75.25%
	v3a	93.07%	87.13%

4.2.3. Impact on MAP

After that, we analyzed the differences in the effects that the various transformation positions had on MAP. Using the UCF-101 dataset, we conduct untargeted attacks against the LRCN model and then calculate the change in MAP that occurs as a result of the attack process. As shown in Figure 4, the MAP and SA of our method remain stable as the number of iterations increases, and they begin to gradually decrease after reaching a certain level of increase. On the other hand, transforming both the perturbation and the original image during the iteration results in a continuous increase of MAP, which is unacceptable for the concealment of adversarial examples. The adversarial examples that are produced by using the accumulative method all have significant perturbations, as can be seen in Figure 5.

Combining the above experimental results and theoretical analysis in Section 3, we can have the following observations:

- The generation process of adversarial examples is the primary focus of existing methods for conducting video adversarial attacks; as a result, the robustness of video adversarial examples has not been investigated. Because video data contain more complex temporal and spatial structures than image data, even relatively minor transformations carried out on the video frames can have a significant effect on the semantic information of the adversarial perturbations. This renders previously developed methods of attack less applicable to the real-world setting in which they are intended to be used.
- Performing transformation during the generation process of video adversarial examples is a practicable solution for enhancing the robustness of adversarial examples. However, repeated transformations can make attacks more difficult or even fail. At the same time, it will also lead to weaker concealment.
- The resilience of video adversarial instances was considerably improved by v3a thanks to the once-transformation that was performed before the attack and the transformation of adversarial perturbations that was performed throughout each iteration. In the meantime, v3a can be effectively integrated with other video adversarial attack methods, which in turn increases the rate of successful attacks.

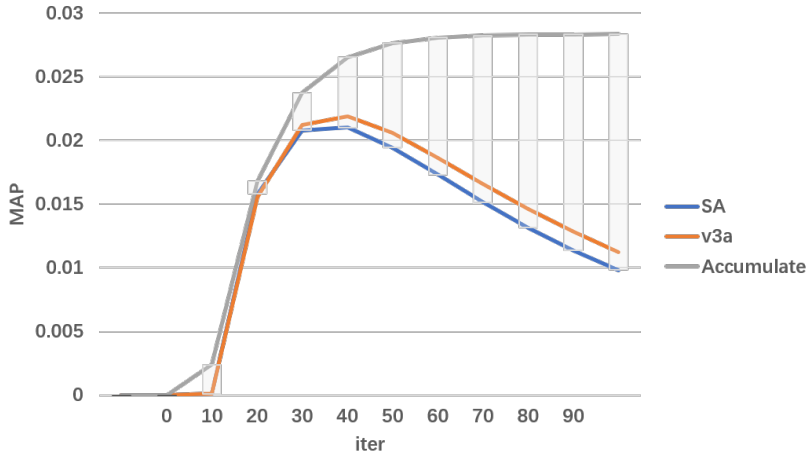

Figure 4. The trend of MAP with iterations on UCF-101 toward the LRCN model.

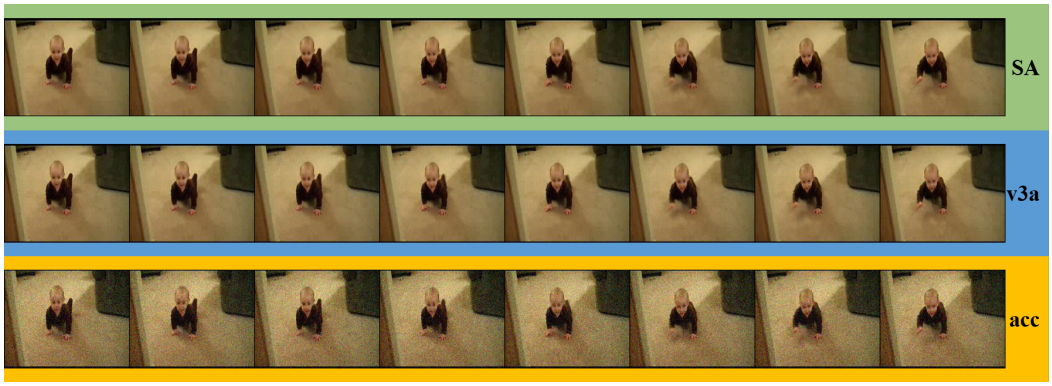

Figure 5. Example of adversarial examples generated by SA, v3a, and the accumulative method, respectively.

5. Discussion

Q: Gaussian noise, video shape transformation, and video color transformation are the three distinct varieties of video alteration that are taken into consideration in version 3a. Defenders, on the other hand, have the ability to develop one-of-a-kind transformations that are inaccessible to attackers. Will the presence of unknown transforms reduce the efficacy of v3a?

A: There may be many different algorithms behind the various video enhancements and transformations. The influence of various transformations, however, may be simply broken down into a rise or a reduction in the value of a given pixel when viewed from the perspective of a single pixel. Because of this, v3a is capable of being transferred across several transformations.

6. Conclusions

In this paper, we proposed an adversarial video attack method that is robust and based on video augmentation. Our method generates adversarial examples that are resistant to environmental transformations such as Gaussian noise, random shape variation, and random color variation. Furthermore, the concealment of adversarial examples is guaranteed by only performing transformations to perturbations, so our method effectively generates examples that are resistant to environmental transformations. Experiments carried out on the UCF-101 video dataset, which is widely utilized, have shown that our method achieves a success rate that is 31.97% higher than that of the existing attack method when it comes to carrying out white-box attacks. However, in comparison to the digital environment, the noise and transitions that may be introduced by the transmission over the air may be

quite different. As a result, the most important work that needs to be done in the future is research into the effect that the physical environment has on video examples and applying the method to real-world attacks.

Author Contributions: Conceptualization, M.Y. and Y.X.; methodology, M.Y. and Y.X.; software, M.Y.; validation, Y.X. and T.H.; formal analysis, Y.X.; writing—original draft preparation, M.Y.; writing review and editing, X.L.; visualization, T.H.; supervision, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 62102379).

Conflicts of Interest: The authors declare no conflict of interest.

Notations and Definitions

All notations and definitions used in our paper:

x	The original input video
δ	The distortion to the original audio
F	The target video classification model
y	The classification result
$\hat{y}$	The target label for targeted attack
α, β	The parameters for controlling perturbations
M	The key frame mask for sparse adversarial attack
L	The loss function
θ	The parameters of the target model
$\hat{g}$	The estimated gradient
N	The number of directions sampled in the gradient estimation process
μ	The smoothing parameter for gradient estimation
ϕ	The directions obtained by random sampling from a Gaussian distribution
s	The step size for the perturbation update
d	The distance evaluation function of the boundary attack
k	The iteration count of the boundary attack
ϵ	The smoothing parameter for the boundary attack
T	The transform function, which performs three different transformations
$\gamma_1, \gamma_2, \gamma_3$	The weights of three transformations

References

1. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
2. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
3. Liu, X.; Du, X.; Zhang, X.; Zhu, Q.; Wang, H.; Guizani, M. Adversarial Samples on Android Malware Detection Systems for IoT Systems. *Sensors* **2019**, *19*, 974.
4. Ding, K.; Liu, X.; Niu, W.; Hu, T.; Wang, Y.; Zhang, X. A low-query black-box adversarial attack based on transferability. *Knowledge-Based Systems* **2021**, *226*, 107102.
5. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 2096–2030.
6. Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv* **2017**, arXiv:1705.07204.
7. Shafahi, A.; Najibi, M.; Xu, Z.; Dickerson, J.; Davis, L.S.; Goldstein, T. Universal adversarial training. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 5636–5643.
8. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), Paris, France, 29–30 April 2017; pp. 39–57.
9. Liu, X.; Hu, T.; Ding, K.; Bai, Y.; Niu, W.; Lu, J. A black-box attack on neural networks based on swarm evolutionary algorithm. In Proceedings of the Australasian Conference on Information Security and Privacy, Perth, WA, Australia, 30 November–2 December 2020; pp. 268–284.
10. Su, J.; Vargas, D.V.; Sakurai, K. One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput.* **2019**, *23*, 828—841.
11. Chen, P.Y.; Sharma, Y.; Zhang, H.; Yi, J.; Hsieh, C.J. EAD: Elastic-net attacks to deep neural networks via adversarial examples. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LO, USA, 2–7 February 2018.

12. Carlini, N.; Wagner, D. Audio adversarial examples: Targeted attacks on speech-to-text. In Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 24 May 2018; pp. 1–7.
13. Kreuk, F.; Adi, Y.; Cisse, M.; Keshet, J. Fooling end-to-end speaker verification with adversarial examples. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 1962–1966.
14. Yuan, X.; Chen, Y.; Zhao, Y.; Long, Y.; Liu, X.; Chen, K.; Zhang, S.; Huang, H.; Wang, X.; Gunter, C.A. Commandersong: A systematic approach for practical adversarial voice recognition. In Proceedings of the 27th USENIX Security Symposium (USENIX Security 18), Montreal, QC, Canada, 10–14 August 2018; pp. 49–64.
15. Qin, Y.; Carlini, N.; Goodfellow, I.; Cottrell, G.; Raffel, C. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition. *arXiv* **2019**, arXiv:1903.10346.
16. Liu, X.; Wan, K.; Ding, Y.; Zhang, X.; Zhu, Q. Weighted-sampling audio adversarial example attack. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 4908–4915.
17. Cheng, Z.; Lu, R.; Wang, Z.; Zhang, H.; Chen, B.; Meng, Z.; Yuan, X. BIRNAT: Bidirectional recurrent neural networks with adversarial training for video snapshot compressive imaging. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 258–275.
18. Zajac, M.; Zołna, K.; Rostamzadeh, N.; Pinheiro, P.O. Adversarial framing for image and video classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 10077–10078.
19. Wei, Z.; Chen, J.; Wu, Z.; Jiang, Y.G. Cross-Modal Transferable Adversarial Attacks from Images to Videos. *arXiv* **2021**, arXiv:2112.05379.
20. Wei, X.; Guo, Y.; Li, B. Black-box adversarial attacks by manipulating image attributes. *Information Sciences* **2021**, *550*, 285–296.
21. Wei, X.; Zhu, J.; Yuan, S.; Su, H. Sparse Adversarial Perturbations for Videos. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8973–8980.
22. Wei, Z.; Chen, J.; Wei, X.; Jiang, L.; Chua, T.S.; Zhou, F.; Jiang, Y.G. Heuristic black-box adversarial attacks on video recognition models. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12338–12345.
23. Xu, Y.; Liu, X.; Yin, M.; Hu, T.; Ding, K. Sparse Adversarial Attack For Video Via Gradient-Based Keyframe Selection. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 2874–2878.
24. Jiang, L.; Ma, X.; Chen, S.; Bailey, J.; Jiang, Y.G. Black-Box Adversarial Attacks on Video Recognition Models. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 864–872.
25. Yan, H.; Wei, X.; Li, B. Sparse black-box video attack with reinforcement learning. *arXiv* **2020**, arXiv:2001.03754.
26. Li, S.; Aich, A.; Zhu, S.; Asif, S.; Song, C.; Roy-Chowdhury, A.; Krishnamurthy, S. Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations. *Adv. Neural Inf. Processing Syst.* **2021**, *34*, 2085–2096.
27. Zhang, H.; Zhu, L.; Zhu, Y.; Yang, Y. Motion-excited sampler: Video adversarial attack with sparked prior. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 240–256.
28. Luo, B.; Liu, Y.; Wei, L.; Xu, Q. Towards Imperceptible and Robust Adversarial Example Attacks Against Neural Networks. In Proceedings of the AAAI, New Orleans, LO, USA, 2–7 February 2018; pp. 1652–1659.
29. Athalye, A.; Engstrom, L.; Ilyas, A.; Kwok, K. Synthesizing Robust Adversarial Examples. In Proceedings of the ICML, Stockholm, Sweden, 10–15 July 2018; pp. 284–293.
30. Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; Song, D. Robust Physical-World Attacks on Deep Learning Visual Classification. In Proceedings of the CVPR. Computer Vision Foundation/IEEE Computer Society, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1625–1634.
31. Brendel, W.; Rauber, J.; Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv* **2017**, arXiv:1712.04248.
32. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
33. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
34. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.