Defending Against Adversarial Videos

Shao-Yuan Lo Vishal M. Patel

Dept. of Electrical and Computer Engineering,
Johns Hopkins University
{sylo, vpatel36}@jhu.edu

Abstract. Adversarial examples of deep neural networks have been actively investigated on image-based classification, segmentation and detection tasks. However, adversarial robustness of video models still lacks exploration. While several studies have proposed how to generate adversarial videos, only a handful of approaches pertaining to the defense strategies have been published in the literature. Furthermore, these defense methods are limited to a single perturbation type and often fail to provide robustness to l_p-bounded attacks and physically realizable attacks simultaneously. In this paper, we propose one of the first defense solutions against multiple adversarial video types for video classification. The proposed approach performs adversarial training with multiple types of video adversaries using independent batch normalizations (BNs), and recognizes different adversaries by an adversarial video detector. During inference, a switch module sends an input to a proper batch normalization branch according to the detected attack type. Compared to conventional adversarial training, our method exhibits stronger robustness to multiple and even unknown adversarial videos and provides higher classification accuracy.

Keywords: Adversarial video, adversarial robustness, physically realizable attack, multiple perturbatios.

1 Introduction

Recent advances in deep learning have led deep neural networks (DNNs) to perform outstandingly well in many computer vision problems [5, 9, 10], including tasks such as video classification [4, 6, 8]. However, researchers have shown that DNNs are easily misled when presented by adversarial examples [7, 26]. This is commonly done by adding carefully crafted perturbations to the input data. Various approaches have also been proposed in the literature to defend against such adversarial attacks [13, 15, 17, 19, 21, 35, 36]. Among them, adversarial training [15, 17, 36] is shown to provide stronger robustness especially to the more challenging white-box and adaptive attacks [1]. As a result, adversarial training has been used as the foundation for more advanced defense techniques. Most recent research in this area has focused on static images. Generating adversarial examples and defense methods for videos is relatively less explored. Although a few recent works have extended adversarial attacks to videos [14, 16, 29, 30, 38],

Fig. 1. Illustration of the four types of adversarial videos we consider on the UCF101 dataset [24]. Top to bottom: Three selected frames from a video. Left to right: Examples of clean, PGD, ROA, AF, and SPA attacks

we are aware of only two studies so far which delve into detecting or defending against adversarial videos [12, 33]. AdvIT [33] is one of the first adversarial frame detectors based on temporal consistency for videos. However, their approach only detects whether a video has been attacked or not. It does not provide any defense mechanism against the attacked videos. Jia et al. [12] leveraged denoising and frame reconstruction for defense. However, it is not clear how well their defense method works on white-box attacks as it was not reported in [12].

On the other hand, conventional adversarial training usually leads to performance degradation on clean data [15, 36]. Xie et al. [34] indicated that this problem is due to the distribution mismatch between clean and adversarial examples. In order to deal with this issue, they leveraged an auxiliary batch normalization (BN) layer [11] to disentangle the two distributions. In addition, most existing adversarial training techniques are generally tailored to one specific perturbation type, e.g., a certain l_p-norm perturbation [17, 20, 31] or physical attacks [32]. A model trained on a specific attack can improve its robustness to that particular attack but often fails to defend when presented with a sample that is perturbed by a different type of attack [23]. Tramèr et al. [28] performed multi-perturbation adversarial training by using Avg and Max strategy towards robustness to different types of l_p perturbations. However, this work does not consider how well it performs on clean images and on unknown attack types. In a real-world application, the input data could be clean (i.e., unattacked), adversarial, or even attacked with a novel attack type that the network has never seen before.

In this paper, we propose one of the first and novel defense solutions for defending against adversarial videos for video classification in the white-box setting, while also considering the accuracy on clean samples as well as the robustness to multiple known and unknown perturbations. Specifically, we consider four of the most significant types of attacks: projective gradient descent (PGD) [17], rectangular occlusion attack (ROA) [32], adversarial framing (AF) [38], and the proposed salt-and-pepper attack (SPA). Fig. 1 gives an illustration of these attacks on video frames. PGD and ROA are originally designed to attack images. We first extend them to videos by perturbing each frame and unveil that video classification models are also vulnerable to these attacks. SPA is a new video attack we design, which looks like salt-and-pepper noise. PGD and SPA belong to the l_p-bounded attack group while ROA and AF belong to the physically realizable attack group. We select one from each group as the known attack type (PGD and ROA) and leave the others as the unknown attack type (AF and SPA), where only the known attacks are used for adversarial training. We aim to defend against all of these attacks while retaining the performance on clean samples simultaneously.

First, we demonstrate that training models on a specific attack can gain robustness to that attack and sometimes to another attack in the same group, but typically cannot defend against the attacks in another group. Training models on multiple attacks together (*multi-perturbation training*) improves *multi-perturbation robustness*, yet accuracy on clean samples is sacrificed. This is mainly due to the distribution mismatch among clean and different types of adversarial examples. We assume that the attacks in the same group have a relatively similar distribution. Therefore, inspired by [34, 37], we employ three BNs in a model: for the clean, l_p-norm, and physically realizable attack examples, individually. Models are trained on these three types of data with their corresponding BN branches. Meanwhile, we create an adversarial video detector to recognize the adversarial group of input video. During inference, a switch module sends the input to a BN branch according to its group. Compared to the conventional adversarial training and multi-perturbation training, experimental results show that the proposed method achieves stronger robustness to multiple, more diverse and even unknown perturbations, while retaining higher accuracy on clean samples.

Our main contributions are summarized as follows:

- We propose a novel defense solution based on a multi-BN structure with a new training scheme and an adversarial video detector. To the best of our knowledge, this is the first defense method against white-box adversarial videos.
- Furthermore, this is the first work to defend against l_p-norm, physically realizable attacks as well as unknown attacks simultaneously.
- The proposed method achieves both stronger multi-perturbation robustness and better performance on clean samples than conventional adversarial training and multi-perturbation training.

2 Related Work

In this section, we review some prior works on adversarial videos, adversarial training, and physically realizable attacks.

Adversarial Videos. Most of the existing literature on adversarial attacks and defense are based on static images. There are only a few works that address attacks and defense techniques for videos. Wei et al. [30] are the first ones to explore adversarial examples in videos. They found that perturbations propagate through video frames in the CNN+RNN based video classifier [6], and thus propose a temporally sparse attack. Li et al. [16] generated video attacks by a generative model. Zajac et al. [38] developed an attack that keeps frames unchanged and just attaches an adversarial frames on the border of each video frame. Jiang et al. [14] introduced V-BAD for black-box video attacks.

Currently, we are aware of only two studies for detecting or defending against video attacks. Xiao et al. [33] proposed AdvIT based on temporal consistency to detect adversarial frames within a video. However, their approach only detects whether a video has been attacked or not. It does not provide any defense mechanism against the attacked videos. Jia et al. [12] presented a similar detector, along with a temporal defense and a spatial defense. The temporal defense reconstructs perturbed frames with adjacent clean frames. And the spatial defense uses and denoises the reconstructed frames to mitigate the effect of adversarial perturbations. However, their approach is only evaluated on the black-box attack setting. It is not clear how well their defense method works on white-box attacks [1] as it was not reported in [12]. In this paper, we provide the first defense solution against white-box adversarial videos.

Adversarial Training. Adversarial training is currently considered the most effective defense approach against adversarial perturbations, particularly for the white-box attacks. Goodfellow et al. [7] first proposed this strategy. They trained DNNs with both clean images and adversarial images to improve adversarial robustness. Madry et al. [17] viewed adversarial training from a perspective of min-max optimization, and trained models with solely adversarial images. It has held great promise and has been widely used as a benchmark. Xie et al. [36] developed a feature denoising block, which increases the network capability of handling adversarial training. Xie et al. [33, 37] also demonstrated that proper normalization management is paramount for enhancing the robustness and to even improve the model performance. Tramèr et al. [28] investigated adversarial robustness to multiple perturbations, including l_∞, l_2, l_1, and rotation-translation attacks, and provided the Max and the Avg adversarial training schemes. Nevertheless, this study does not consider potential unknown attack types and clean images into consideration. Our proposed defense is based on adversarial training and manages normalization to enhance the robustness to multiple, more diverse and even unknown perturbations, while retaining higher accuracy on clean images simultaneously.

Physical Attacks. The physical attack is a class of adversarial attacks that can be implemented in the physical space. The physically realizable attack refers to the digital representation of physical attacks. Such attacks fool DNNs by modifying physical objects being photographed. Sharif et al. [22] generated printable perturbations inside eyeglass frames to attack face recognition systems. Brown et al. [3] created an adversarial patch that can be put next to a real-world object making that object be misclassified. Thys et al. [27] further extended the adversarial patch to fooling human detectors. Wu et al. [32] proposed DOA, a pioneering defense against physically realizable attacks. DOA performs adversarial training with rectangular occlusion attack (ROA), which places an adversarial rectangular sticker on an image, improving physical robustness. However, it fails to resist l_p-bounded attacks. Our work extends ROA/DOA to videos and shows robustness to both l_p-bounded and physically realizable attacks.

3 Multiple Adversarial Video Types

For our investigation, we construct four types of video attacks: l_∞-norm PGD [17], ROA [32], AF [38], and a new SPA attack (see Fig. 1). Among them, l_∞-norm PGD and SPA (l_0-norm) belong to the l_p-bounded attacks, and ROA and AF belong to the physically realizable attacks. In our experiments, we set PGD and ROA as the known attack types used for training, while AF and SPA are used as unknown attack types that are only used during inference. We aim to defend against multiple adversarial video types, including l_p-bounded and physically realizable attacks as well as known and unknown attacks. All of these attacks are set to untargeted since the untargeted attack is considered more difficult to defend against than the targeted attack.

Projective Gradient Descent. PGD attack is defined as

$$x^{t+1} = \Pi_{x+\mathbb{S}} \left(x^t + \alpha \, sgn(\bigtriangledown_x L(x, y; \theta)) \right) \tag{1}$$

where x is a data sample, y is the ground-truth label, θ is model parameters, L is the training loss, and $\mathbb{S}$ denotes the set of allowed perturmations. The perturbation size ϵ is described as $\parallel x^T - x \parallel_p \leq \epsilon$, where T denotes the maximum interation number. PGD is a powerful multi-step variant of the Fast Gradient Sign Method (FGSM) [7]. It has become one of the most important benchmarks in the current adversarial example research [32, 36, 37].

We extend the l_∞-norm PGD from images to videos by taking the gradient descent with respect to an entire input video. We set the perturbation size $\epsilon = 4$, the number of iterations $n = 5$, and step size $\alpha = 1$, for both adversarial training and evaluation.

Rectangular Occlusion Attack. ROA attack introduces l_∞-norm PGD inside a fixed-size and fixed location rectangle on an image. The size is pre-defined and the location is searched with respect to the highest loss. We extend ROA to

videos, in which each frame is perturbed by a rectangle. To save computations, we skip the location search step. Instead, we randomly assign the rectangle location for each video frame and then apply PGD on it. We set the rectangle size equal to 30×30 with PGD setting $\epsilon = 255$, $n = 5$, and $\alpha = 1$, for both adversarial training and evaluation.

Adversarial Framing. AF attack adds adversarial framings on the border of each video frame while most of the frame pixels are not modified. It is originally designed as a universal attack. To save computations, we perform a non-universal version by first fixing the framing location and then applying PGD inside it. We set the framing width equal to 5 with PGD setting $\epsilon = 255$, $n = 5$, and $\alpha = 1$, for both adversarial training and evaluation.

Salt-and-pepper Attack. Inspired by the one-pixel attack [25], we design a new video attack type. For computation saving, instead of using differential evolution, we randomly select a pre-defined number of pixels on each video frame, then apply PGD on those pixels. We consider it as a kind of l_0-norm attack because the number of adversarial pixels is bounded. We name this new attack *salt-and pepper attack (SPA)*, as the perturbations look like salt-and-pepper noise. We set the number of adversarial pixels on each frame equal to 100 with PGD setting $\epsilon = 255$, $n = 5$, and $\alpha = 1$, for both adversarial training and evaluation.

4 Defense Against Adversarial Videos

In real-world applications, the input data could be clean (i.e. unattacked), adversarial, or even attacked with a novel attack that the network has never seen before. As a result, it is important to design a defense solution that can resist to multiple known and unknown perturbations while retaining the performance on clean samples. In this section, we first revisit adversarial training and multi-perturbation training schemes, and then introduce the proposed method which is based on multiple BNs and an adversarial video detector. Fig. 2 gives an overview of the proposed approach.

4.1 Adversarial Training

To begin with, we recall the objective function for training a DNN model

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{D}} \left[L(x, y; \theta) \right], \tag{2}$$

where x is a clean training sample with ground-truth label y in the training set $\mathbb{D}$, θ is model parameters, and L denotes the training loss. Madry's adversarial

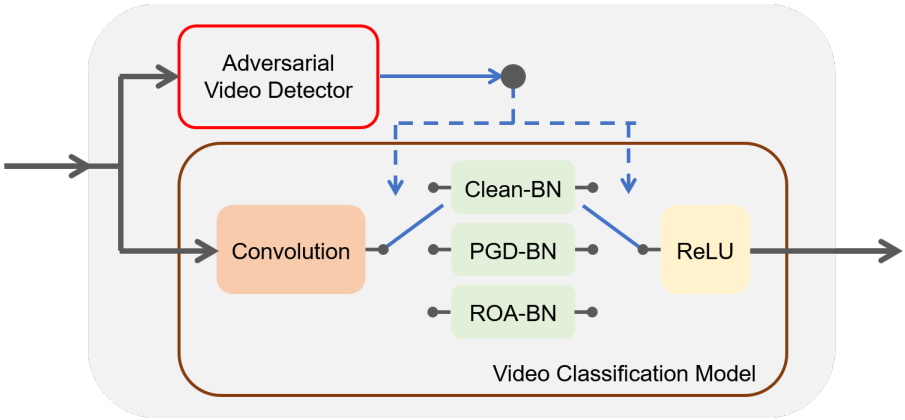

Fig. 2. An overview of the proposed framework for defense

training [17] applies the min-max optimization and trains models exclusively on adversarial examples

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{D}} \left[\max_{\delta\in\mathbb{S}} L(x+\delta, y; \theta) \right], \tag{3}$$

where δ denotes an adversarial perturbation in the perturbation set $\mathbb{S}$. AdaProp [34] aims to improve the performance on clean samples and trains the model with a mixture of clean data and adversarial examples as follows

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{D}} \left[L(x, y; \theta) + \max_{\delta\in\mathbb{S}} L(x+\delta, y; \theta) \right]. \tag{4}$$

Regarding multi-perturbation robustness, Tramèr et al. [28] introduce two adversarial training strategies: Avg strategy and Max strategy. Avg strategy trains on all types of adversarial examples simultaneously and optimizes these adversarial losses together as follows

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{D}} \left[\sum_{i=1}^{N} \max_{\delta_i\in\mathbb{S}_i} L(x+\delta_i, y; \theta) \right], \tag{5}$$

where N is the number of perturbation types. Max strategy considers the worse-case attack. It trains on the strongest adversarial example that obtains the maximum loss among all types of attacks

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{D}} \left[\max_{\delta_k\in\mathbb{S}_k} L(x+\delta_k, y; \theta) \right], \tag{6}$$

where

$$k = \arg\max_{i\in[1,N]} \left[\max_{\delta_i\in\mathbb{S}_i} L(x+\delta_i, y; \theta) \right], \tag{7}$$

which denotes the strongest type of attacks. With this background, in what follows, we elaborate on our proposed training scheme which is based on a multiple BN structure.

4.2 Multi-BN Structure

The adversarial training on a single perturbation type is generally weakly robust to the other types of attacks. On the other hand, most current state-of-the-art computer vision DNNs contain BNs [11] in their architecture to normalize input features, which improves the performance [8, 10]. However, because of the different distributions among multiple perturbation types, BNs suffer from the distribution mismatch when multi-perturbation training is conducted, and thus fails to gain promising multi-perturbation robustness.

To address this problem, we deploy multiple BN branches into each BN layer of the target model and keep the rest of the parts unchanged, i.e., still a single network [34, 37]. Clean data and each perturbation type we use for training are assigned an individual BN branch. Each BN branch is only responsible to estimate one or similar distributions, and thus can properly disentangle multiple distributions. At each mini-batch training step, each perturbation attacks the model through its assigned BN branch. Then we send these adversarial and clean mini-batches to their corresponding BN branch and compute the loss. The objective function is defined as follows

$$\theta^* = \arg\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\mathbb{D}} \left[L(x, y; \theta^c, \theta_0^b) + \sum_{i=1}^{N} \max_{\delta_i \in \mathbb{S}_i} L(x + \delta_i, y; \theta^c, \theta_i^b) \right], \quad (8)$$

where θ^c is convolution parameters, θ_i^b is the BN parameters of the ith data type, and $\theta = \theta^c + \sum_{i=0}^{N} \theta_i^b$ denotes all the model parameters.

In practical scenarios, DNNs have to be able to provide robustness against unknown attacks. An exhaustive investigation is too expensive; instead, we can summarize different attack types into several groups based on their distributions, then build a BN branch for each group. In our case, we adopt three BN branches: for clean data, l_p-norm attacks, and physically realizable attacks (see Fig. 2). Typically, adversarial training on a strong and well-designed attack has better robustness [17]. According to our analysis in Sec.5.2, PGD and ROA are a better l_p-norm attack and a better physically realizable attack, respectively for adversarial training among the four types of attacks we consider. Therefore, we train models on clean, PGD, and ROA examples using the 3-BN structure with Eq. (8), where N=2.

4.3 Adversarial Video Detector

During inference, the input data must be fed to the corresponding BN branch. Hence, an automatic switch mechanism is needed. To this end, we propose to use a video classification-based adversarial video detector which not only identifies

whether an input video is clean or adversarial, but also recognizes the attack type. Then a switch module sends the input to a proper BN branch according to the detection results (see Fig. 2). The adversarial video detector is a N+1 class classifier, where N is the number of attack types used for training. We employ a 3D CNN architecture [8] and train it end-to-end.

For our case, the detector is trained by N=2 on clean, PGD, and ROA examples to recognize clean data, l_p-bounded attacks, and physically realizable attacks. For the unknown attacks, the detector would classify them into the most similar group on the basis of Softmax probabilities. Our experiments in Sec.5.2 demonstrate that AF videos and SPA videos are mostly classified into the l_p-bounded group and the physically realizable group, respectively, which is tailored to our system design though these examples are not used for training. We are aware that vigorously classifying an unknown class as one of the known classes is not a proper solution. However, since our ultimate purpose is defense, this is exactly what we want in our specific case. Applying an open-set recognition algorithm [2, 18] to make the detector classify the unknown attck types is left for future work.

5 Experiments

We evaluate the robustness of our method to the four considered perturbation types as well as its performance on clean data. In this section, we first describe our experimental setup. Next, we report the evaluation results of the proposed method. Finally, an ablation study is conducted to analyze different components of the proposed method.

5.1 Experimental Setup

Implementation Details. We choose 3D ResNext-101 and 3D Wide ResNet-50 [8] as our target models, as they are two of the most top-performing 3D CNNs for video classification. We adopt the pre-trained models from [8] and conduct adversarial training to fine-tune them in end-to-end. SGD optimizer is used for all models and experiments with learning rate 0.001, momentum 0.9, and weight decay $1e^{-5}$. For our adversarial video detector, we employ the lightweight 3D ResNet-50 and train it with the same training protocol as mentioned above.

Dataset. We train and evaluate our method on UCF101 [24], a widely used video dataset in action recognition. It consists of 13,320 videos with 101 action classes. The dataset splits 9,537 videos for training and 3,783 videos for testing. We use the whole training set to train the target model and the whole test set for evaluation. For training our adversarial video detector, 3,733 videos are selected from the training set to construct a self-made 11,199-video set containing clean, PGD, ROA classes with an equal number of videos. The frame dimensinos are 112×112. Following [30], we uniformly sample each video into 40 frames.

5.2 Evaluation and Ablation Study

Effectiveness of 3-BN Framework. We evaluate both the robustness and the performance of the proposed 3-BN training scheme on 3D ResNext-101. We compare it with conventional adversarial training [17] as well as the Avg and Max multi-perturbation training strategies [28]. Different from [28], we take accuracy of clean samples into consideration. For a fair comparison, we adjust the Avg and Max strategies by involving clean data in training. In particular, we add the clean data loss $L(x, y; \theta)$ into the expectation of objective functions Eq. (5) and Eq. (6).

Table 1 shows that models trained on a specific attack have the best robustness to that attack compared to the other single attack trained model. PGD_{AT} and ROA_{AT} also yield high robustness to another attacks in their own group, showing better generalization. However, all of them almost fail to defend against the attacks from other groups. Avg_{MPT} and Max_{MPT} that use multi-perturbation training achieve better multi-perturbation robustness, yet their performance on clean data is not promising owing to the distribution mismatch. Oddly, Max_{MPT} obtains higher accuracy than PGD_{AT} under the PGD attack. Note that Max_{MPT} considers the worse-case attack and thus favors a stronger attack, but a detailed analysis is left for future analysis.

We can see that the proposed 3BNs_{MPT} substantially outperforms Avg_{MPT} and Max_{MPT} on clean data as well as on general attacks. The evaluation on 3D Wide ResNet-50 is reported in Table 2. Again, our 3BNs_{MPT} achieves significantly better performance and robustness.

Table 1. Evaluation results on 3D ResNext-101. *Clean* model is trained on only clean data. *AT* and *MAT* denote adversarial training and multi-perturbation training, respectively. PGD_{AT}, ROA_{AT}, AF_{AT} and SPA_{AT} are models trained on a single attack, while Avg_{MPT} and Max_{MPT} are trained on multiple attacks using Avg and Max strategies. 3BNs_{MPT} is our 3-BN structure with *MPT*. Acc_{adv} and Acc_{all} are the mean accuracies (%) over four attacks and over all the five input types. The best results are in bold, and the best results among *MPT* methods are underlined.

Model	Clean	PGD	ROA	AF	SPA	Acc_{adv}	Acc_{all}
Clean	**80.3**	2.2	1.9	11.8	9.8	6.4	21.2
PGD_{AT}	68.4	36.1	5.9	8.5	**60.8**	27.8	35.9
ROA_{AT}	74.9	2.6	56.4	68.2	19.3	36.6	44.3
AF_{AT}	65.6	1.3	8.3	**77.4**	12.7	24.9	33.1
SPA_{AT}	77.8	4.2	3.7	32.1	56.4	24.1	34.8
Avg_{MPT}	70.7	36.2	53.2	66.1	52.1	51.9	55.7
Max_{MPT}	69.3	**44.0**	47.8	62.3	53.4	51.9	55.4
3BNs_{MPT}	<u>76.9</u>	34.9	**59.6**	<u>70.6</u>	<u>58.5</u>	**55.9**	**60.1**

Table 2. Evaluation results on 3D Wide ResNet-50. *Clean* model is trained with only clean data. *MAT* denotes multi-perturbation training. Avg_{MPT} and Max_{MPT} are models trained on multiple attacks using Avg and Max strategies. 3BNs_{MPT} is our 3-BN structure with *MPT*. Acc_{adv} and Acc_{all} are the mean accuracies (%) over four attacks and over all the five input types. The best results are in bold, and the best results among *MPT* methods are underlined.

Model	Clean	PGD	ROA	AF	SPA	Acc_{adv}	Acc_{all}
Clean	**77.3**	2.5	1.2	6.0	8.9	4.7	19.2
Avg_{MPT}	67.9	**34.3**	52.9	62.4	52.0	50.4	53.9
Max_{MPT}	67.0	33.8	41.7	57.4	53.7	46.7	50.7
3BNs_{MPT}	<u>74.3</u>	30.2	**55.7**	**68.9**	**55.6**	**52.6**	**56.9**

Table 3. Performance of each BN branch on the five input types. Clean_{BN}, PGD_{BN} and ROA_{BN} are the clean, PGD and ROA BN branches in the 3-BN structure, respectively.

BN Branch	Clean	PGD	ROA	AF	SPA
Clean_{BN}	**76.9**	13.9	43.2	59.0	28.4
PGD_{BN}	68.3	**34.9**	12.6	30.5	**58.5**
ROA_{BN}	76.6	15.3	**59.6**	**70.6**	31.4

Ablation on Differnt BN Branches. In the previous section, the attacker generates perturbations through the BN branch corresponding to its type. During inference, the input is also sent to that BN branch accordingly. In this section, we investigate the cases that perturbations are produced on a BN branch which is different from the group of the used attack type, and the input is sent to that BN branch during inference.

As can be seen in Table 3, each BN branch performs the best on the input type which they are trained on. Moreover, for the unknown attacks, PGD_{BN} is the most robust to SPA, and ROA_{BN} is the most robust to AF. This result is consistent to our hypothesis that feeding an unknown adversarial example to the BN branch of the same or the most similar group can enjoy the best benefit. Our grouping is based on the observation that PGD and SPA have similar distributions (l_p-bounded attacks), and ROA and AF have similar distributions (physically realizable attacks) is also confirmed.

Table 4. Inf Target

Attack		PGD			ROA	
Inf \ Target	Clean_{BN}	PGD_{BN}	ROA_{BN}	Clean_{BN}	PGD_{BN}	ROA_{BN}
Clean_{BN}	13.9	41.3	20.5	43.2	44.1	54.3
PGD_{BN}	**61.7**	34.9	**61.0**	33.5	12.6	38.0
ROA_{BN}	26.1	41.2	15.3	**66.2**	**65.6**	**59.6**
Attack		AF			SPA	
Inf \ Target	Clean_{BN}	PGD_{BN}	ROA_{BN}	Clean_{BN}	PGD_{BN}	ROA_{BN}
Clean_{BN}	59.0	62.1	66.2	28.4	47.8	34.7
PGD_{BN}	42.9	30.5	43.3	**65.6**	**58.5**	**65.6**
ROA_{BN}	**72.5**	**71.1**	**70.6**	35.1	47.9	31.4

Ablation on Target BN and Inference BN. Next, we further delve into the cases that an adversarial example is made inference on a BN branch (inference BN) different from the BN branch that is used to generate the adversarial example (target BN). In other words, we consider the cases that the target BN and inference BN are different.

The results shown in Table 4 are consistent with Table 3 in general, in which PGD_{BN} has the strongest robustness to l_p-bounded attacks, and ROA_{BN} has the strongest robustness to physically realizable attacks. PGD attack is an exception. When the target BN is PGD_{BN}, inference PGD_{BN} performs the worst.

In addition, we observe that for any specific inference BN, it is more robust to the adversarial examples generated on another BN branch, i.e., target BN and inference BN are different. In such a case, the attack is not a rigorous white-box attack, so we treat it as a kind of gray-box attack, in which the attacker does not know which BN branch would adversarial examples will be sent to during inference. This result unveils that the attacks cannot perfectly transfer to other BN branches even though the rest of the model parameters are shared in the same network. Meanwhile, the robustness of the proposed 3-BN framework to this gray-box attack is demonstrated, and thus its black-box robustness is also optimistic.

Equipping Automatic Switch. Now we equip the 3-BN structure with the adversarial video detector and switch module, making a complete system. We consider two white-box attack cases to evaluate the composite:

– Case A: The attacker is able to access the target model directly and is fully aware of the switch codes. In other words, the attacker can attack through any target BN he likes. To compare with Table 1, we set the target BN corresponding to the attack type and set inference BN to the same as the target BN.

Table 5. Evaluation results on 3-BN framework equipped with adversarial video detector. Case Z corresponds to the case when adversarial video detector has a perfect prediction corresponding to Table 1. The numbers in parentheses are the detection accuracies (%) of the adversarial video detector.

Case	Clean	PGD	ROA	AF	SPA	Acc_{all}
Z	76.9 *(100.0)*	34.9 *(100.0)*	59.6 *(100.0)*	70.6 *(100.0)*	58.5 *(100.0)*	60.1 *(100.0)*
A	76.4 (92.0)	36.6 (95.6)	60.5 (100.0)	70.7 (99.8)	59.1 (98.3)	60.7 (97.1)
B	76.4 (92.0)	76.2 (8.7)	65.0 (0.0)	62.8 (0.0)	67.3 (41.1)	69.5 (28.4)

– Case B: The composite is treated as a complete single system. The attacker can access the entire system but cannot access the inner component. In other words, the attacker generates perturbations by the gradients of the entire end-to-end system. The target BN and inference BN depend on the prediction of the adversarial video detector.

We report the results in Table 5. Case Z is assuming the adversarial video detector has a perfect prediction, which is actually the same case as in Table 1 that we send inputs to BN branches manually. We put it here for comparison. The accuracy of the detector in Case A is equivalent to the standard evolution on it since the generated perturbations do not depend on the detector. We can see that our detector achieves good performance (97.2%). For the target model in Case A, the clean clean accuracy is slightly lower, the reason is that some of the clean videos are mis-detected to the PGD class. Recall our results in Table 3, PGD_{BN} performs worse on clean data than Clean_{BN}. On the other hand, some PGD data are mis-detected to the clean class, which causes a gray-box attack, so the accuracy becomes higher. The performance differences in the rest of the attacks are due to the error from the randomization of perturbation initialization.

Moving to Case B, the clean data inference accuracy and detection accuracy of Case B is the same as Case A since the difference between these two cases is only on perturbation generation. On the other hand, in Case B, we observe that the detection accuracies are significantly lower for all perturbations (even 0% in some cases). Most PGD, ROA and SPA examples are mis-detected to the clean class. AF examples are mis-detected to both clean and l_p-norm classes. Meanwhile, the inference accuracies are higher for all perturbations except AF. The reason is that it is more difficult to generate adversarial examples by using the gradients obtained from the whole system. The produced perturbations are too weak to be detected by the adversarial video detector and cause just a weak threat against the target model.

Fig. 3. Illustration of the adversarial videos generated under Case B. Top to bottom: Three selected frames from a video. Left to right: Examples of clean, PGD, ROA, AF, and SPA attacks

Fig. 3 illustrates some examples of the adversarial videos generated under Case B. As can be seen from this figure, they look different from Fig. 1. In this case, PGD examples are almost identical to clean data, so PGD inference accuracy is very close to clean accuracy. ROA examples are with gray rectangles nearly unchanged from the initialization, and thus mostly assigned to Clean_{BN} by the detector. The gray rectangles occlude objects. Therefore, sometimes they can also fool the target model, making ROA inference accuracy lower than clean accuracy. AF examples remain perturbations on the framings, but the perturbations are not strong enough to be detected as the physically realizable attack class. The lower AF inference accuracy is due to the worse robustness of Clean_{BN} and PGD_{BN} to the AF attacks (see Table 3 and Table 4) though the perturbations are weaker in this case. SPA examples are basically with just black holes, and they are weaker adversaries. Hence, SPA inference accuracy is higher. Based on the above observations, the composite of the 3-BN structure and the adversarial video detector makes it difficult for attackers to produce adversarial videos in Case B. This indicates that the proposed method can further enhance the robustness in certain cases in a different way.

6 Conclusions

The adversarial robustness of video networks still lacks exploration. Very few defense methods in the video domain have been introduced. In this paper, we propose one of the first defense solutions against adversarial videos based on a

multi-BN structure with a new training scheme and an adversarial video detector. The proposed method can defend against lp-norm attacks and physically realizable attacks simultaneously in an unknown attack setting. Compared to conventional adversarial training and multi-perturbation training, it provides more promising performance on clean data and multi-perturbation robustness, simultaneously. We conduct an extensive ablation study to analyze different cases. Furthermore, reducing the expense of multi-perturbation training is also a challenge. These issues are left for future work. Problems on adversarial machine learning in videos and multi-perturbation robustness are worth further exploration.

References

1. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: International Conference on Machine Learning (2018)
2. Bendale, A., Boult, T.E.: Towards open set deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
3. Brown, T., Mane, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. In: Conference and Workshop on Neural Information Processing Systems Workshop (2017)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: IEEE conference on Computer Vision and Pattern Recognition (2017)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In: IEEE transactions on pattern analysis and machine intelligence (2017)
6. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015)
8. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
9. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: IEEE International Conference on Computer Vision (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
11. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (2015)
12. Jia, X., Wei, X., Cao, X.: Identifying and resisting adversarial videos using temporal consistency. arXiv preprint arXiv:1909.04837 (2019)
13. Jia, X., Wei, X., Cao, X., Foroosh, H.: Comdefend: An efficient image compression model to defend adversarial examples. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)

14. Jiang, L., Ma, X., Chen, S., Bailey, J., Jiang, Y.G.: Black-box adversarial attacks on video recognition models. In: ACM International Conference on Multimedia (2019)
15. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. In: International Conference on Learning Representations (2017)
16. Li, S., Neupane, A., Paul, S., Song, C., Krishnamurthy, S.V., Roy-Chowdhury, A.K., Swami, A.: Stealthy adversarial perturbations against real-time video classification systems. In: Network and Distributed System Security Symposium (2019)
17. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018)
18. Oza, P., Patel, V.M.: C2ae: Class conditioned auto-encoder for open-set recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
19. Raff, E., Sylvester, J., Forsyth, S., McLean, M.: Barrage of random transforms for adversarially robust defense. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
20. Raghunathan, A., Steinhardt, J., Liang, P.: Certified defenses against adversarial examples. In: International Conference on Learning Representations (2018)
21. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In: International Conference on Learning Representations (2018)
22. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: ACM Conference on Computer and Communications Security (2016)
23. Sharma, Y., Chen, P.Y.: Attacking the madry defense model with l_1-based adversarial examples. In: International Conference on Learning Representations Workshop (2018)
24. Soomro, K., Zamir, A.R., Shah, M., Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
25. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. In: IEEE Transactions on Evolutionary Computation (2019)
26. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (2014)
27. Thys, S., Van Ranst, W., Goedemé, T.: Fooling automated surveillance cameras: adversarial patches to attack person detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (2019)
28. Tramèr, F., Boneh, D.: Adversarial training and robustness for multiple perturbations. In: Conference on Neural Information Processing Systems (2019)
29. Wei, X., Liang, S., Chen, N., Cao, X.: Transferable adversarial attacks for image and video object detection. In: International Joint Conferences on Artificial Intelligence (2019)
30. Wei, X., Zhu, J., Yuan, S., Su, H.: Sparse adversarial perturbations for videos. In: AAAI Conference on Artificial Intelligence (2019)
31. Wong, E., Kolter, J.Z.: Provable defenses against adversarial examples via the convex outer adversarial polytope. In: International Conference on Machine Learning (2018)
32. Wu, T., Tong, L., Vorobeychik, Y.: Defending against physically realizable attacks on image classification. In: International Conference on Learning Representations (2020)

33. Xiao, C., Deng, R., Li, B., Lee, T., Edwards, B., Yi, J., Song, D.X., Liu, M., Molloy, I.: Advit: Adversarial frames identifier based on temporal consistency in videos. In: IEEE International Conference on Computer Vision (2019)
34. Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A., Le, Q.V.: Adversarial examples improve image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
35. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. In: International Conference on Learning Representations (2018)
36. Xie, C., Wu, Y., van der Maaten, L., Yuille, A., He, K.: Feature denoising for improving adversarial robustness. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
37. Xie, C., Yuille, A.: Intriguing properties of adversarial training at scale. In: International Conference on Learning Representations (2020)
38. Zajac, M., Zołna, K., Rostamzadeh, N., Pinheiro, P.O.: Adversarial framing for image and video classification. In: AAAI Conference on Artificial Intelligence (2019)

A Evaluation on Stronger Attacks

In this section, we test the proposed 3-BN framework (3BNs$_{MPT}$) on stronger adversarial attacks, and compare it with Avg (Avg$_{MPT}$) and Max (Max$_{MPT}$) strategies. These models are identical to those in Table 1, in which they are trained on clean data, PGD examples with perturbation size $\epsilon = 4$, the number of iterations $n = 5$, and step size $\alpha = 1$, and ROA examples with rectangle size 30×30, $\epsilon = 255$, $n = 5$, and $\alpha = 1$ (see Sec. 3).

We establish two sets of stronger attacks. In the first set, we increase ϵ from 4 to 8 for PGD, rectangle size from 30×30 to 45×45 for ROA, framing width from 5 to 10 for AF, and the number of adversarial pixels on each frame from 100 to 200 for SPA. In the second set, we further increase ϵ to 16 for PGD, rectangle size to 60×60 for ROA, framing width to 20 for AF, and the number of adversarial pixels on each frame to 400 for SPA.

As reported in Table 6 and Table 7, the proposed 3BNs$_{MPT}$ still achieves much better multi-perturbation robustness than Avg$_{MPT}$ and Max$_{MPT}$ even when tested on stronger attacks.

Table 6. Evaluation results under the first set of stronger attacks. PGD† is PGD attack with $\epsilon = 8$, ROA† is ROA attack with rectangle size 45×45, AF† is AF attack with framing width 10, and SPA† is SPA attack with 200 adversarial pixels on each frame. The best results are in bold.

Model	PGD†	ROA†	AF†	SPA†	Acc$_{all}$
Clean	0.7	0.3	3.8	2.6	1.8
Avg$_{MPT}$	13.8	33.7	63.9	40.7	38.0
Max$_{MPT}$	**24.2**	27.5	58.8	45.2	38.9
3BNs$_{MPT}$	16.8	**37.4**	**67.8**	**48.7**	**42.7**

Table 7. Evaluation results under the second set of stronger attacks. PGD‡ is PGD attack with $\epsilon = 16$, ROA‡ is ROA attack with rectangle size 60×60, AF‡ is AF attack with framing width 20, and SPA‡ is SPA attack with 400 adversarial pixels on each frame. The best results are in bold.

Model	PGD‡	ROA‡	AF‡	SPA‡	Acc_{adv}
Clean	0.0	0.1	0.6	0.3	0.3
Avg_{MPT}	2.5	**21.2**	52.4	22.9	24.7
Max_{MPT}	5.2	13.7	40.0	**31.1**	22.5
3BNs_{MPT}	**6.7**	20.2	**53.1**	27.5	**26.9**

B Equipping Automatic Switch for 3D Wide ResNet-50

We also evaluate the performance of the 3-BN structure with the adversarial video detector and switch module on 3D Wide ResNet-50. Again, the detector architecture is based on 3D ResNet-50 and is trained on clean videos as well as the PGD and ROA adversarial videos generated on the target model. We report the results in Table 8. These results are consistent with the case of 3D ResNeXt-101 in Table 5.

Table 8. Evaluation results of 3-BN framework equiped with adversarial video detector on 3D Wide ResNet-50. Case Z corresponds to the case when the adversarial video detector has a perfect prediction corresponding to Table 2. The numbers in parentheses are the detection accuracies (%) of the adversarial video detector.

Case	Clean	PGD	ROA	AF	SPA	Acc_{all}
Z	74.3 *(100.0)*	30.2 *(100.0)*	55.7 *(100.0)*	68.9 *(100.0)*	55.6 *(100.0)*	56.9 *(100.0)*
A	73.2 (92.0)	32.2 (94.4)	55.9 (100.0)	68.4 (82.2)	55.9 (98.3)	57.1 (93.4)
B	73.2 (92.0)	73.3 (8.9)	62.4 (0.0)	61.5 (0.0)	63.8 (51.5)	66.8 (30.5)