Analysis and Extensions of Adversarial Training for Video Classification

Kaleab A. Kinfu and René Vidal
Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, MD, USA

Abstract

Adversarial training (AT) is a simple yet effective defense against adversarial attacks to image classification systems, which is based on augmenting the training set with attacks that maximize the loss. However, the effectiveness of AT as a defense for video classification has not been thoroughly studied. Our first contribution is to show that generating optimal attacks for video requires carefully tuning the attack parameters, especially the step size. Notably, we show that the optimal step size varies linearly with the attack budget. Our second contribution is to show that using a smaller (sub-optimal) attack budget at training time leads to a more robust performance at test time. Based on these findings, we propose three defenses against attacks with variable attack budgets. The first one, Adaptive AT, is a technique where the attack budget is drawn from a distribution that is adapted as training iterations proceed. The second, Curriculum AT, is a technique where the attack budget is increased as training iterations proceed. The third, Generative AT, further couples AT with a denoising generative adversarial network to boost robust performance. Experiments on the UCF101 dataset demonstrate that the proposed methods improve adversarial robustness against multiple attack types.

1. Introduction

Deep neural networks (DNNs) have led to significant advances in many computer vision tasks, including image classification, object detection, and image segmentation. However, a key challenge to DNNs is their vulnerability to adversarial attacks [35], i.e., small perturbations to the input data that fool the classifier. This has motivated the development of various defense mechanisms against such attacks. Among them, Adversarial Training (AT) [10,24,42], a technique that extends the training set with examples that maximize the loss, shows state-of-the-art adversarial robustness. Nevertheless, the robust performance against adversarial attacks is still far from perfect as there is a significant difference between clean accuracy and adversarial accuracy.

In this work, we are interested in studying the adversarial robustness of DNNs for classifying actions in video data. Action classification aims to predict a single action label for a video and is one of the main tasks in video understanding. Numerous DNN-based methods have been developed for this task, *e.g.*, [5, 9, 12, 20], which exploit the spatio-temporal information from consecutive frames in the video.

Unlike image attacks, video attacks consider not only spatial but also temporal information. For example, flickering attacks [27] generate a flickering temporal perturbation whereas other attacks, such as the frame saliency attacks [16], generate temporally sparse perturbations. Moreover, AT has been found to be difficult to successfully train a robust model on large-scale problems [21], thus it is a natural extension to study the effectiveness of AT for videos.

Although adversarial attacks in the case of videos are not as investigated as in the image domain, a few attack generating methods have been recently proposed [16, 17, 27, 40]. Similarly, defenses in the video domain are less explored. Xia *et al*. [41] proposed adversarial frame detection from videos leveraging temporal consistency by applying optical flow estimation to consecutive frames. However, this technique can only tell if the video is adversarial or not. Lo *et al*. [23] proposed using AT with multiple independent batch normalization layers. However, the experiments are limited to weak attacks, and the robust performance against stronger attacks has not yet been reported.

Paper Contributions. In this paper, we study the effectiveness of AT for video classification. Since AT is based on generating attacks that maximize the loss during training, we begin by understanding how to generate such optimal attacks. We observe that generating optimal attacks requires carefully tuning the attacker parameters, such as the attack budget, the step size, and the number of steps. In particular, we show that there is an optimal choice for the step size that changes linearly with the attack budget. Next, we focus on understanding whether optimal attacks at training time lead to better defenses at test time. Surprisingly, the answer to this question is no. Specifically, we show that using attacks with a smaller (sub-optimal) attack budget during training leads to a more robust performance at test time. Based on these findings, we propose three defenses, adaptive AT, curriculum AT and generative AT, which use sub-optimal attacks with an adaptive attack budget during training. In summary, this paper makes the following contributions:

- We show that generating optimal attacks is very sensitive to the choice of the step size and requires careful tuning. We also show that the optimal step size varies for different attack budgets and that their relationship is linear.
- Contrary to the common belief that the key to solving the min-max problem in AT is to find optimal adversarial examples [15], we empirically demonstrate that AT with an optimal attack does not lead to a strong defense. Rather, we show that AT with a smaller attack budget leads to a better defense. We also show that AT with a higher attack budget does not necessarily lead to a better defense.
- Since AT with a single perturbation type does not often yield robustness to multiple attack types, we propose AT with an adaptive random sampling of variable attack types and budgets. Moreover, based on our finding that AT with optimal attacks fails, we propose curriculum based adversarial learning with an increasing attack budget as training proceeds. We also propose to combine adversarial AT with a GAN-based perturbation elimination network [18] as a pre-processor to improve robust performance. Finally, we demonstrate the robustness of the proposed methods with experiments on the UCF101 dataset.

2. Related Work

In this section, we briefly review adversarial attack methods related to this work as well as the adversarial training technique that we will study. In addition, we briefly discuss generative defense techniques related to our method.

Adversarial Attacks. Given an image X and a classifier F with parameter θ_F, an additive adversarial attack is a small perturbation δ to the image X which is designed to deliberately fool the classifier, i.e. $F(X+\delta;\theta_F) \neq F(X;\theta_F)$. The perturbation δ is assumed to be sufficiently small so that they are imperceptible to humans. A common assumption is that the perturbations lie within a ball $\mathcal{B}_\epsilon = \{\delta : \|\delta\|_p \leq \epsilon\}$, where ϵ denotes the attack budget and $p \in \{0,1,2,\infty\}$.

Ever since Szegedy *et al.* [35] first observed that DNNs are vulnerable to adversarial attacks, several attack methods have been proposed by the research community. These attack methods can be classified in two categories: white-box and black-box attacks. White-box attacks assume access to the gradient information of the model, while in black-box attacks, the attacker does not have access to the parameters of the model. White-box attacks [3, 10, 24] basically maximize the loss value by iteratively computing a perturbation and adding it to the data, whereas black-box attacks [7, 44] are usually performed by using a different model and transferring adversarial examples to the target model.

In this work, we focus on white-box attacks obtained by maximizing the cross-entropy loss, i.e.

$$\hat{\delta} = \operatorname*{arg\,max}_{\delta \in \mathcal{B}_\epsilon} \mathcal{L}_{ce}(F(X+\delta;\theta_F), y), \tag{1}$$

where y is the label for image X. Goodfellow *et al.* [10] propose the Fast Gradient Sign Method (FGSM), a single step gradient based attack, to compute the perturbation δ:

$$\delta = \epsilon \operatorname{sign}(\nabla_\delta \mathcal{L}_{ce}(F(X;\theta_F), y)). \tag{2}$$

Madry *et al.* [24] extend this by introducing an iterative approach to finding the optimal attack known as Projected Gradient Descent (PGD), which can be formulated as:

$$\delta_{i+1} = \Pi_{\mathcal{B}_\epsilon}\left\{\delta_i + \alpha \operatorname{sign}(\nabla_\delta \mathcal{L}_{ce}(F(X+\delta_i;\theta_F), y))\right\}, \tag{3}$$

where $\Pi_{\mathcal{B}_\epsilon}$ denotes the projection of the perturbation onto the norm ball $\mathcal{B}_\epsilon$, while α is the step size and i is the current step. In this work, we will mostly consider PGD bounded ℓ_∞-based attacks and denote it as PGD$_\infty$.

Adversarial Attacks to Video Data. So far, most adversarial attack studies have focused on the image domain, leaving attacks for videos less explored. However, there are few attack techniques for videos. *Masked PGD attacks* apply the PGD$_\infty$ attack to a patch of a frame given a random top-left position and a patch ratio r. The patch height and width are computed by multiplying the frame height and width by $\sqrt{r}$. *Frame Border attacks* apply the PGD$_\infty$ attack to the border of a frame instead of a random patch. The thickness of the border b given a ratio r is calculated as a function of the image size. Beyond ℓ_∞ attacks, *sparse adversarial attacks* [40] fool action classification networks by using $\ell_{2,1}$ optimization to compute the perturbations.

Frame saliency attacks [16] extend iterative gradient-based attacks to prioritize which parts of the video frames should be perturbed based on saliency scores. The baseline variant is the *one-shot attack*, which only involves one iteration of the attack. Here, the loss is computed with respect to each input frame. The other variant is an *iterative-saliency* attack that generates adversarial examples by iteratively perturbing some frames of the video one at a time in order of decreasing saliency score until it fools the network.

Flickering attacks [27] generate a flickering temporal perturbation that is imperceptible to human observers by adding a uniform offset to the C color channels of the pixels in every frame of a video. The perturbation is computed by optimizing the following objective function:

$$\hat{\delta} = \operatorname*{arg\,min}_{\delta} \lambda \sum_j \beta_j D_j(\delta) + \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_{cw}(F(X+\delta;\theta_F), y), \tag{4}$$

where $\mathcal{L}_{cw}$ denotes the C&W loss [3], λ controls the importance of the adversarial attack relative to the regularization terms $D_j(\cdot)$, which control the imperceptibility of the attack. The parameter β_j weights the relative importance of each regularization term. Here, the regularization terms are thickness regularization that enforces the perturbation to be small in all color channels and roughness regularization that penalizes temporal changes of the perturbation pattern.

Adversarial Defenses. The discovery of vulnerabilities in DNNs has raised security concerns and attracted significant attention from the research community. Accordingly, various defense methods have been proposed [14]. Papernot *et al*. [26] propose a distillation defense that uses knowledge from the distillation process to reduce the magnitude of network gradient which could successfully mitigate perturbations computed by FGSM. Guo *et al*. [11] use a transformation technique that replaces input patches with clean patches to remove the perturbations. Dhillon *et al*. [8] propose applying dropout to each layer in order to prune a random subset of activations with a weighted distribution. While such methods do not require retraining the model, Athalye *et al*. [1] show that they suffer from the obfuscated gradients problem, which gives a false sense of robustness.

Adversarial training (AT) [10, 35] is based on solving a min-max optimization to achieve robustness via retraining the model with adversarial examples. Shaham *et al*. [31] formalize AT as the following robust optimization problem:

$$\min_{\theta_F} \mathop{\mathbb{E}}_{(X,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{B}_\epsilon} \mathcal{L}_{ce}(F(X+\delta;\theta_F),y) \right], \tag{5}$$

where (X, y) is the training data sampled from the data distribution $\mathcal{D}$. The inner maximization problem aims to generate an optimal adversarial example for each (X, y), while the outer minimization problem trains the model to be robust to such adversarial examples.

AT using the PGD attack to solve the inner maximization problem is a widely used technique that yields state-of-the-art performance in many image-based tasks [37]. However, Kurakin *et al*. [21] show that AT is difficult at ImageNet scale and Sharma and Chen [32] show that training using ℓ_∞ adversarial examples yields limited robustness.

Generative Adversarial Defenses. Song *et al*. [33] empirically show that most adversarial examples lie in low probability regions of the training data distribution, thus proposing a generative method that tries to project adversarial examples back onto the clean data manifold prior to classification. This technique is based on the idea of purification and it uses PixelCNN [39] to approximately find the highest probability example constrained to be within an ϵ-ball of a perturbed image. Similarly, Samangouei *et al*. [29] propose to use a Generative Adversarial Network (GAN), instead of PixelCNN, to project adversarial examples back onto the manifold of the generator. On the other hand, She *et al*. [18] propose an approach called Adversarial Perturbation Elimination with GAN (APE-GAN) that allows for training a GAN to systematically clean the perturbations from the adversarial examples thus learning a manifold mapping from adversarial data to clean data. Lin *et al*. [22] propose a method known as Dual Manifold Training that uses perturbations in the latent space of StyleGAN [19] as well as in the image space in order to adversarially train a model with the aim of defending against ℓ_p and non-ℓ_p attacks. However, this requires a pre-constructed On-Manifold dataset, which is too expensive for video data.

3. Empirical Analysis of Adversarial Training with PGD Attacks for Action Classification

In this section, we present an empirical analysis of adversarial training (AT) with PDG attacks for action classification in the UCF101 dataset. We first show that a naïve application of PGD to action classification fails to produce optimal attacks, i.e., attacks that maximize the loss. Specifically, we show that generating optimal attacks requires carefully tuning the parameters of PGD, especially the step-size. We also show that different attack budgets require a different step size in order for PGD to find an optimal attack, and that the relationship between the optimal step size and the attack budget is linear. As a consequence, we obtain a practical procedure for obtaining optimal attacks for action classification via suitable tuning of the PGD parameters. Next, we show that optimal attacks during training do not lead to a strong defense. Specifically, we show that choosing an attack budget at training that is smaller than the attack budget at testing leads to a more robust performance at testing. This motivates us to propose various modifications to AT and generative defenses, which will be discussed in Sec. 4.

3.1. Optimal PGD Attacks via Parameter Tuning

Let us first recall the definition of an optimal attack for video classification. Let $X \in \mathbb{R}^{T\times H\times W\times C}$ denote a video with T frames, where H and W denote, respectively, the height and width of each frame, and C denotes the number of color channels. Let $y \in \{1, \dots, K\}$ denote the label for the video. In this case, the perturbation is also a video denoted as $\delta = [\delta_1, \delta_2, \dots, \delta_T] \in \mathbb{R}^{T\times W\times H\times C}$, where δ_t denotes the perturbation to the t^{th} frame x_t of the video X.

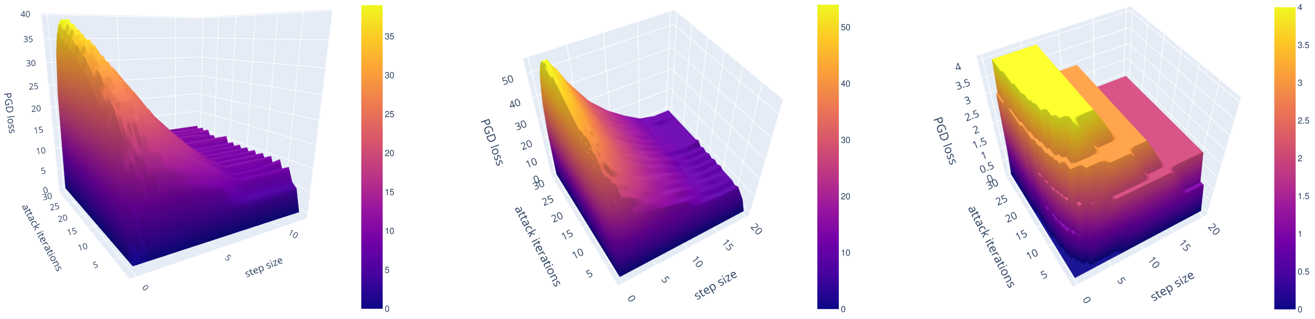
Given an action classification model F with parameters θ_F, and an attack budget $\epsilon > 0$, an optimal adversarial attack is computed by finding a perturbation δ that maximizes, *e.g*., the cross-entropy loss function $\mathcal{L}_{ce}$, *i.e*.:

$$\hat{\delta} = \arg\max_{\delta\in\mathcal{B}_\epsilon} \mathcal{L}_{ce}(F(X+\delta;\theta_F),y), \tag{6}$$

where $\mathcal{B}_\epsilon = \{\delta : \|\delta\|_p \leq \epsilon\}$ is the ℓ_p-norm ball with radius ϵ, and $p \in \{0, 1, 2, \infty\}$. The Projected Gradient Descent (PGD) method solves the above optimization problem using the following iterative algorithm for $i = 1, \dots, M$,

$$\delta_{i+1} = \Pi_{\mathcal{B}_\epsilon}\left\{\delta_i + \alpha\,\text{sign}(\nabla_\delta \mathcal{L}_{ce}(F(X+\delta_i;\theta_F),y))\right\}, \tag{7}$$

where $\Pi_{\mathcal{B}_\epsilon}$ is the projection of the perturbation onto the ball $\mathcal{B}_\epsilon$, α is the step size and M is the number of iterations.



(a) Benign ($\epsilon = 4$)

(b) Benign ($\epsilon = 8$)

(c) AT ($\epsilon = 8$)

Figure 1. (a) Adversarial loss surface of a benign model trained on clean data when the input is perturbed by PGD$_\infty$ attack with a budget $\epsilon = 4$, (b) Adversarial loss surface of the same benign model perturbed by PGD$_\infty$ attack with a budget $\epsilon = 8$. (c) Adversarial loss surface of the same model after trained with adversarial examples for 10 epochs with a budget $\epsilon = 8$. The x-axis is the step size α; the y-axis is the number of steps M, and the z-axis is the adversarial loss value. From (a) and (b), observe that generating optimal attacks require tuning the step size in a thin region. Specifically, a small step size, i.e. below some threshold, generates a very weak attack, while increasing the step size after the optimal value also reduces the attack strength. Moreover, the adversarial loss becomes a flat when $\alpha \geq 2\epsilon$. From (b) and (c) we can see that the adversarial loss decreases significantly and the step size region that can generate the optimal attack becomes wider.

In order to evaluate the effects of the attack parameters α and M in generating an optimal perturbation for a given ϵ, we visualize the adversarial loss of a benign model trained on clean data. We adopt a 3D CNN architecture based on residual networks (ResNets) [13], namely 3D ResNeXt-101 [12], as a backbone for the action recognition task. We adopt the training strategy and configuration of parameters from [20]. The network was first pre-trained on Kinetics-700 [4], a large-scale video dataset with more than 650,000 videos covering 700 categories. This step is important as such networks require an enormous amount of data. The Kinetics-700 pretrained network was then fine-tuned on the UCF101 dataset [34], a widely used dataset in action recognition. The UCF101 dataset consists of 13,320 videos with 101 action classes providing a large diversity of actions, variations in lighting conditions, camera motion, viewpoint, and partial occlusion. The dataset is split into 9,537 videos for training and 3,783 videos for testing.

Figure 1 shows the adversarial loss of the benign model for a PGD$_\infty$ attack as a function of the step size α and the number of iterations M for a fixed value of $\epsilon = 4$ (1a) and $\epsilon = 8$ (1b). As it can be seen from the plot, there is a very small range of values of α that will generate the optimal attack. Indeed, a step size below the threshold will generate a very weak attack, while the loss becomes flat if a step size above 2ϵ is used. Note that the optimal step size changes with the attack budget, since $\hat{\alpha} \approx 1$ for $\epsilon = 4$ and $\hat{\alpha} \approx 1.8$ for $\epsilon = 8$, hence α must be properly tuned.

We further investigated the relationship between the optimal step size $\hat{\alpha}$ and the attack budget ϵ by computing the loss surfaces for multiple values of ϵ in the range $[1, 12]$ and finding the optimal α for each ϵ. As we can see from Figure 2, there is a linear relationship between the optimal step size and attack budget, which is given by $\hat{\alpha} \approx 0.2(\epsilon + 1)$.

Next, we investigate the effect of AT on the adversarial loss. Figure 1b shows the adversarial loss surface of the benign model for $\epsilon = 8$ and Figure 1c shows the adversarial loss surface of an adversarially trained model for 10 epochs. We observe that, as the iterations proceed, (1) the adversarial loss surface becomes approximately piece-wise constant, (2) there is a wider range of values for the step size that yield an optimal attack, (3) the optimal value of the step size for the benign model is still optimal for the adversarially trained model, and (4) the adversarial loss value at the optimum decreases significantly. Therefore, for a fixed ϵ, as the iterations of AT proceed, there is no need to change the value of α and finding effective attacks becomes harder.

3.2. Adversarial Training with Optimal Attacks Gives Suboptimal Performance at Test Time

Adversarial training is predicated on the idea of training with optimal attacks. For example, Huang *et al.* [15] claim that the key to solving the min-max problem of AT is to find optimal adversarial examples. The results from the previous section show us how to tune the parameters of PGD to obtain optimal attacks for video classification. In this subsection we wish to verify whether AT with optimal attacks gives a strong defense. Moreover, we wish to understand if the attack budget at training time should be chosen to be the same as the attack budget at test time.

We validate the effectiveness of AT with the PGD$_\infty$ attack for different attack budgets $\epsilon \in \{4, 8, 12\}$ and step sizes $\alpha \in \{\hat{\alpha}, \hat{\alpha}/2, \epsilon\}$ using the top-1 test accuracy on the UCF101 dataset. The results of the benign and adversarially trained models for different attack budgets at test time are shown in Table 1. As expected, the benign model per-

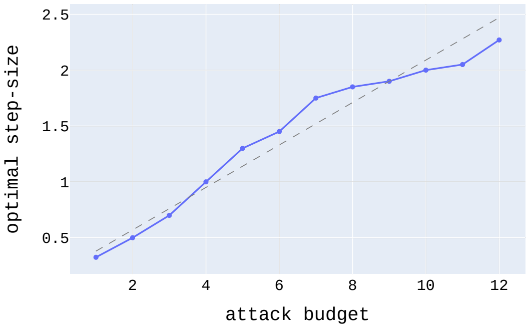


Figure 2. **Optimal step size vs attack budget:** The relationship between the optimal step size $\hat{\alpha}$ and the attack budget ϵ is well approximated by linear regression (dash line) as $\hat{\alpha} \approx 0.2(\epsilon + 1)$.

forms well on clean data ($\epsilon = 0$), but its accuracy deteriorates quickly from 92.3% to around 2% when evaluated with adversarial examples. On the other hand, AT with optimal attacks, AT($\epsilon = 8, \alpha = \hat{\alpha}$), has lower standard accuracy (28.8%) and higher robust accuracy (9.39%-20.5%) than the benign model, but it generally underperforms AT with sub-optimal attacks, AT($\epsilon = 8, \alpha \neq \hat{\alpha}$), in terms of both standard and robust accuracy.

Another interesting observation is that the step size in sub-optimal regions controls the trade-off between standard accuracy and robustness. Loosely speaking, a smaller step size ($\alpha = \hat{\alpha}/2$) helps preserve standard accuracy to some extent but suffers from poor robustness, while increasing the step size ($\alpha = \epsilon$) improves robustness but reduces the standard accuracy. The results also show that using a stronger attack budget during training ($\epsilon = 12$) reduces both standard and robust accuracy, while using a weaker attack budget ($\epsilon = 4$) improves standard accuracy as well as robustness against weaker attack ($\epsilon = 4$).

Overall, the results thus far suggest that when the attack budget at test time is known, we should use AT with the same attack budget at training time and a suboptimal choice for the step size. However, when the attack budget at test time is unknown, we need a mechanism for selecting the attack budget and step size at training time. This motivates the extensions of AT that we discuss next.

4. Extensions of Adversarial Training

Based on our findings from the previous section, in this section we propose three defenses against attacks with variable attack budgets. The first one, Adaptive AT, is a method where the attack budget is drawn from a distribution that is adapted as training iterations proceed. The second one, Curriculum AT, is a technique where the attack budget is increased as training iterations proceed. The third one, Generative AT, further couples AT with a denoising generative network to boost robust performance. The overall architecture of the proposed methods is illustrated in Figure 3.

Table 1. Top-1 test accuracy on the UCF101 dataset of both benign and adversarially trained models with PGD$_\infty$ for different attack budgets and step sizes. All attacks at test time are computed using PGD$_\infty$ with a fixed attack budget and the optimal step size.

Model	Evaluation Attack				
	$\epsilon = 0$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 12$	$\epsilon = 15$
Benign	**92.3**	2.65	2.38	2.11	1.89
AT($\epsilon = 8, \alpha = \hat{\alpha}$)	28.8	20.5	13.9	10.6	9.39
AT($\epsilon = 8, \alpha = \frac{\hat{\alpha}}{2}$)	*72.9*	*40.0*	19.5	11.0	8.22
AT($\epsilon = 8, \alpha = \epsilon$)	64.1	39.9	**23.3**	**13.6**	*10.6*
AT($\epsilon = 4, \alpha = \epsilon$)	70.1	**40.5**	*19.9*	12.8	10.4
AT($\epsilon = 12, \alpha = \epsilon$)	34.3	24.8	17.8	*13.5*	**12.1**

4.1. Adaptive Adversarial Training (AAT)

As it can be seen from the AT framework (see Eq. 5), the attack budget is predetermined for a fixed attack type. While this can provide robustness to that particular attack type and budget, it does not guarantee robustness against different attacks at test time [36]. For example, Schott *et al.* [30] show that increasing robustness to one attack type can decrease robustness to other attacks. Besides, the results in Table 1 show that a model trained with an attack budget ϵ almost always outperforms models trained with a different budget when evaluated on attacks with the same budget ϵ.

Here, we propose a technique to train the model with adversarial examples generated from a discrete set of attack types $\mathcal{T}$, *e.g.* $\mathcal{T}$ = {Frame Saliency, Frame Border, PGD Patch, Flickering}, or a discrete set of B attack budgets $\mathcal{E} = \{\epsilon_1, \ldots, \epsilon_B\}$ to yield simultaneous robustness against multiple attacks at a test time. A naïve approach to handle this could be to sample the attack type or budget from a uniform distribution on $\mathcal{T}$ or $\mathcal{E}$ and generate an attack with that type or budget. However, the resulting model could be more robust to some attacks and less robust to others. To address this issue, we propose to adapt the distribution of attack types or budgets depending on the performance of the current model for different attacks. For example, we can calculate the probability of attack budget $\epsilon_i \in \mathcal{E}$ as

$$\mathbb{P}(\epsilon_i) := \frac{\sum_{n=1}^{N} \mathcal{L}_{ce}(F(X_n + \delta_n^i; \theta_F), y_n)}{\sum_{j=1}^{B} \sum_{n=1}^{N} \mathcal{L}_{ce}(F(X_n + \delta_n^j; \theta_F), y_n)}, \tag{8}$$

where δ_n^i is an attack for (X_n, y_n) with budget ϵ_i. As adapting this probability at each epoch can be costly, we update it only every ξ epochs. Algorithm 1 summarizes the proposed AAT algorithm with adaptive sampling of multiple attack budgets. Multiple attack types can be handled similarly.

4.2. Curriculum Adversarial Training (CAT)

The framework of AT tries to optimize the inner maximization by generating the optimal adversarial examples,

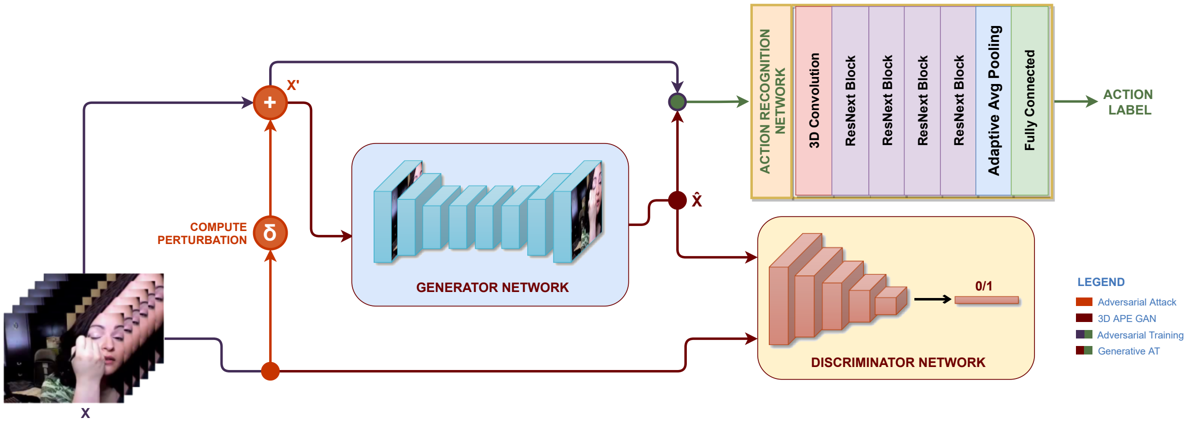


Figure 3. **Overall architecture of the proposed methods –** Given a video X, an adversarial perturbation δ is added to it to generate an adversarial example X', which is then passed to the action classifier model for Adversarial Training (AT). Adaptive AT and Curriculum AT extend AT by changing the attack budget and attack type during training. In the case of 3D APE-GAN, the generator network is trained to eliminate perturbations from X' and reconstruct $\hat{X}$, which is then passed to the benign model, while the discriminator is trained to classify the original video X from reconstructed video $\hat{X}$. Generative AT is an end-to-end integration of 3D APE-GAN and AT.

Algorithm 1 Adversarial training with an adaptive random sampling of variable attack budget

```
Require: set of attack budgets $\mathcal{E}$, set of attack step sizes $\mathcal{A}$
function ADAPTAT
    randomly initialize $\mathbb{P}(\epsilon)$ for $\epsilon \in \mathcal{E}$
    for $k \leftarrow 1$ to $K$ do                                  ▷ number of epochs
        for $n \leftarrow 1$ to $N$ do                              ▷ number of batches
            $\epsilon \sim \mathbb{P}$                               ▷ sample attack budget
            $\alpha = \mathcal{A}(\epsilon)$                         ▷ set step size for $\epsilon$
            $\delta_n \leftarrow$ ATTACKER$(F, X_n, y_n, \epsilon, \alpha)$     ▷ Eq. 7
            $\theta_F \leftarrow \theta_F - \eta \nabla_{\theta_F} \mathcal{L}_{ce}(F(X_n + \delta_n; \theta_F), y_n)$
        end for
        if $k \mod \min(\xi, \frac{K}{|\mathcal{E}|}) \equiv 0$ then
            for $i \leftarrow 1$ to $|\mathcal{E}|$ do
                $\mathbb{P}(\epsilon_i) := \frac{\sum_{n=1}^{N} \mathcal{L}_{ce}(F(X_n + \delta_n^i; \theta_F), y_n)}{\sum_{j=1}^{B} \sum_{n=1}^{N} \mathcal{L}_{ce}(F(X_n + \delta_n^j; \theta_F), y_n)}$
            end for
        end if
    end for
end function
```

however, adversarially robust generalization might hurt the standard generalization [38]. Plus, the optimal adversarial attack does not always lead to better defense, or it might even lead the training to fail and result in a worse generalization in both standard and robust performance [43]. It has also been found that AT overfits the adversarial examples observed [2, 28]. To mitigate all these issues, we propose a curriculum-based AT, wherein the model is trained with adversarial examples generated with an increasing attack budget at every ξ^{th} iteration as the training proceeds. Cai *et al*. [2] propose a similar approach to the image classification task, however, it only increases the number of steps while the attack budget is fixed, and again this does not provide guaranteed robustness for other attack budgets [36].

4.3. Generative Adversarial Training (GAT)

Defense methods based on deep generative models such as GANs have shown promising results in the image domain. However, these techniques lag in video generation and state-of-the-art results are very far from satisfying [6].

In this work, we propose 3D APE-GAN, an extension of APE-GAN [18] that removes perturbations from adversarial videos by projecting them onto a clean data manifold. Specifically, given an adversarial video, the 3D APE-GAN generator G is trained to reconstruct a clean video, while the discriminator D is trained to distinguish the original video from the video produced by the generator. This is achieved by solving a min-max problem of the form:

$$\min_{\theta_G} \max_{\theta_D} \mathop{\mathbb{E}}_{(X,y)\sim\mathcal{D}} \Big[\gamma_1 \mathcal{L}_{mse}(G(X + \hat{\delta}; \theta_G), X) \\ + \gamma_2 \big(\log(D(X; \theta_D)) - \log(D(G(X + \hat{\delta}; \theta_G); \theta_D))\Big], \tag{9}$$

where θ_G denotes parameters of the generator, θ_D denotes parameters of the discriminator, $\mathcal{L}_{mse}$ is the MSE loss

$$\mathcal{L}_{mse} := \frac{1}{TWH} \|X - G(X'; \theta_G)\|_F^2, \tag{10}$$

and $\hat{\delta} := \arg\max_{\delta \in \mathcal{B}_\epsilon} \mathcal{L}_{ce}(F(G(X + \delta; \theta_G); \theta_F), y)$.

Notice that the min-max objective is a weighted combination of the MSE loss and the GAN objective with weights γ_1 and γ_2, respectively. We solve this problem in an alternating minimization fashion, where given θ_G, we find θ_D as

$$\max_{\theta_D} \mathop{\mathbb{E}}_{(X,y)\sim\mathcal{D}} \log(D(X; \theta_D)) - \log(D(G(X + \hat{\delta}; \theta_G); \theta_D)). \tag{11}$$

Then, given θ_D, we find θ_G as

$$\min_{\theta_G} \mathop{\mathbb{E}}_{(X,y)\sim\mathcal{D}} \Big[\gamma_1 \mathcal{L}_{mse}(G(X+\hat{\delta};\theta_G), X) \\ -\gamma_2 \log(D(G(X+\hat{\delta};\theta_G);\theta_D))\Big]. \quad (12)$$

One extension we propose is to train 3D APE-GAN with AT so that the generator can be used as a preprocessing step to eliminate the perturbations before being fed to the video action recognizer. We can first train the 3D APE-GAN to generate clean videos given adversarial videos, freeze the pre-trained generator and adversarially train only the action recognition network F. This can be formulated as:

$$\min_{\theta_F} \mathop{\mathbb{E}}_{(X,y)\sim\mathcal{D}} \Big[\max_{\delta\in\mathcal{B}_\epsilon} \mathcal{L}_{ce}(F(G(X+\delta;\theta_G);\theta_F), y)\Big]. \quad (13)$$

Another extension we propose is to train 3D APE-GAN and the action recognition network end-to-end by optimizing:

$$\min_{\theta_F}\min_{\theta_G}\max_{\theta_D} \mathop{\mathbb{E}}_{(X,y)\sim\mathcal{D}} \Big[\gamma_1 \mathcal{L}_{mse}(G(X+\hat{\delta};\theta_G), X) \\ +\gamma_2\big(\log(D(X;\theta_D)) - \log(D(G(X+\hat{\delta};\theta_G);\theta_D)\big) \quad (14) \\ + \mathcal{L}_{ce}(F(G(X+\hat{\delta};\theta_G);\theta_F), y)\Big].$$

5. Experiments and Results

As discussed before, the results in Table 1 show that a model trained with an attack budget ϵ almost always outperforms models trained with a different budget when evaluated on attacks with the same budget ϵ. This suggests we need to train our model with multiple attack budgets to provide better robustness to different attack at test time. In this section, we evaluate the performance of the proposed AAT, CAT and GAT methods against multiple attack types.

5.1. AT with Variable Attack Budget

The first method we evaluate is AAT, *i.e.*, the proposed adversarial training method with an adaptive random sampling of variable attack budgets, as presented in Algorithm 1. The network is trained with a set of attack budgets $\mathcal{E} = \{0, 2, 4, .., 12\}$ sampled based on the probability distribution computed via Eq. 8 updated every $\xi = 10$ epochs. The second approach we evaluated is the proposed Curriculum AT (CAT) where we increase the attack budget from $\epsilon = 0$ to $\epsilon = 12$ or decrease it from $\epsilon = 12$ to $\epsilon = 0$ as the training proceeds. The overall results are shown in Table 2.

As we can see from the results, AAT improves the robust performance (especially against stronger attacks). However, the standard accuracy reduces by almost 10% when compared to AT. On the other hand, CAT with an increasing attack budget outperforms all the other previous methods in all adversarial attacks while it lessens the standard accuracy by only 3% from AT. That being said, although CAT improves robust accuracy, the overfitting problem [28] is still not fully solved: Figure 4 shows that the generalization gap goes up as the training iterations and attack budget simultaneously increase. This demonstrates that the generalization gap when training with stronger attacks is higher than with weaker attacks. Thus, this suggests that further work is needed to close this gap in order to yield better robustness against optimal attacks at test time.

Table 2. **Adversarial training with variable attack budgets** – Top-1 test accuracy on the UCF101 dataset of both benign and adversarially trained models with PGD$_\infty$ for variable attack budgets and step sizes. All attacks at test time are computed using PGD$_\infty$ with a fixed attack budget and the optimal step size. AAT with attack budget $\epsilon = auto$ represents the adaptive random sampling method of variable attack budgets. CAT with $\epsilon = \uparrow_0^{12}$ indicates curriculum training with an increasing attack budget from $\epsilon = 0$ to $\epsilon = 12$, while CAT with $\epsilon = \downarrow_0^{12}$ indicates a decreasing budget from $\epsilon = 12$ to $\epsilon = 0$.

Model	Evaluation Attack				
	$\epsilon = 0$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 12$	$\epsilon = 15$
Benign	**92.3**	2.65	2.38	2.11	1.89
AT($\epsilon = 8, \alpha = 8$)	*64.1*	*39.9*	23.3	13.6	10.6
AAT($\epsilon = auto, \alpha = \epsilon$)	54.5	38.3	*25.7*	*18.7*	*16.0*
CAT($\epsilon = \downarrow_0^{12}, \alpha = \epsilon$)	47.8	24.4	11.5	6.61	5.40
CAT($\epsilon = \uparrow_0^{12}, \alpha = \epsilon$)	61.1	**42.2**	**29.2**	**21.0**	**17.2**

5.2. Generative Adversarial Training

In this experiment, we evaluate the 3D APE-GAN method and the end-to-end trainable GAT method. The performance of both techniques with the previous best performing technique is presented in Table 3. The results show that the 3D APE-GAN with benign model significantly improves the standard and robust accuracy when compared with previous AT methods. Moreover, GAT further improves the performance in almost all evaluations. However,

Table 3. **3D APE-GAN with AT** – Top-1 test accuracy on the UCF101 dataset of both benign and adversarially trained models with PGD$_\infty$ for different attack budgets and step sizes. All attacks at test time are computed using PGD$_\infty$ with a fixed attack budget and the optimal step size. APE denotes the proposed 3D APE-GAN method using benign action recognition network, while GAT represents the Generative AT, i.e. end-to-end adversarial training of both 3D APE-GAN and action recognition network.

Model	Evaluation Attack				
	$\epsilon = 0$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 12$	$\epsilon = 15$
Benign	**92.3**	2.65	2.38	2.11	1.89
AT($\epsilon = 8, \alpha = 8$)	64.1	39.9	23.3	13.6	10.6
CAT($\epsilon = \uparrow_0^{12}, \alpha = \epsilon$)	61.1	**42.2**	**29.2**	**21.0**	**17.2**
APE($\epsilon = \uparrow_0^{12}, \alpha = \epsilon$)	*81.6*	*69.7*	*59.8*	*54.6*	**54.1**
GAT($\epsilon = 8, \alpha = 8$)	81.1	**72.6**	**63.8**	**56.6**	*51.3*

Table 4. **Robustness to multiple attack types** –Test accuracy on 100 videos of the UCF-101 dataset of benign and adversarially training models with different video attack types. For each attack type, the results presented here are an average top-1 test accuracy of three experiments with varying parameters. The results of each attacks with the different parameters is depicted in Figure 5.

Model	Evaluation Attack					
	No attack	Frame saliency (one-shot)	Frame saliency (iterative)	Masked PGD	Frame Border	Flickering
Benign	**93.0**	0.00	0.00	7.0	0.00	42.3
Video Compression [25]	92.6	48.6	50.0	7.0	10.0	40.3
AAT ($\epsilon = auto, \alpha = \epsilon$)	92.6	*74.4*	74.9	*67.5*	*80.2*	53.1
APE ($\epsilon = auto, \alpha = \epsilon$)	*92.7*	74.0	*76.6*	63.1	63.4	*55.3*
GAT ($\epsilon = auto, \alpha = \epsilon$)	79.3	**85.1**	**86.3**	**81.9**	**86.6**	**58.1**

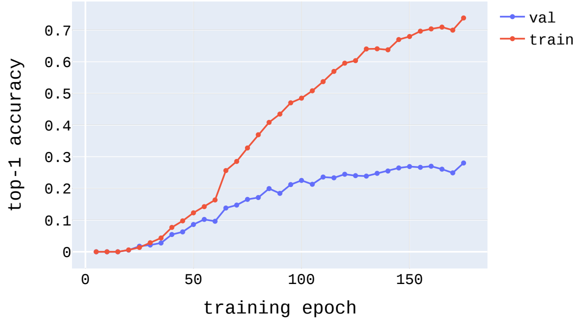


Figure 4. **Generalization gap of curriculum adversarial training –** The generalization gap keeps increasing as the curriculum training proceeds. The x-axis is training iterations and the y-axis is the top-1 accuracy. The red and blue lines indicate the top-1 training and validation accuracy, respectively.

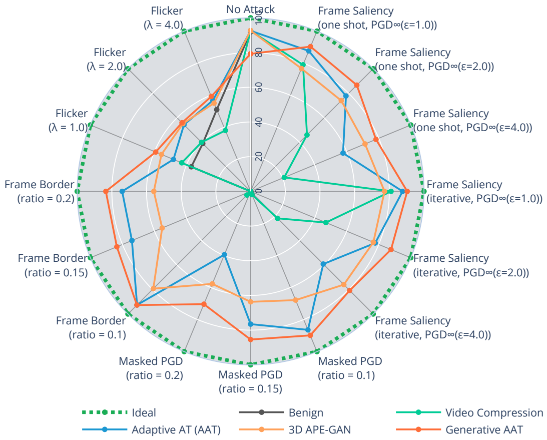


Figure 5. **Robustness to multiple attack types** – robust performance of models trained with different video attack types evaluated on 100 videos of the UCF-101 dataset.

note that the generator is being used as a pre-processing step and thus it is not being attacked.

5.3. AT with Multiple Attack Types

In this experiment, we evaluate the effectiveness of our methods against different attack types. We consider five video attack methods – frame saliency (one-shot), frame saliency (iterative), masked PGD, frame border, and flickering attacks. The baseline method we use is the video compression defense [25]. The methods are trained with adversarial examples generated from an adaptive distribution of four attack methods (except the flickering attack). The robustness performance of the methods to the different video attacks is illustrated in Figure 5. The average performance against each attack type is listed in Table 4. From the results, it is evident that AAT provides robustness to multiple attack types and Generative Adaptive AT boosts the robustness, but reduces the standard accuracy by just above 10%.

6. Conclusions

In this paper, we showed that generating optimal attacks for video requires carefully tuning the attack parameters, especially the step size. We also showed that the optimal step size varies linearly with the attack budget. We demonstrated that using a sub-optimal attack budget at training time leads to more robust performance at testing, and that the generalization gap rises as the attack budget increases during testing. Based on these findings, we proposed three defenses against attacks. Adaptive AT and Curriculum AT extend AT by using variable attack budgets as the training iterations proceed. This is achieved by adapting the distribution of attack budgets to the performance and increasing the attack budget, respectively. Generative AT integrates AT with a denoising generative network to further improve robust performance. Our proposed approaches significantly increase the robustness to multiple attack budgets and attack types.

Acknowledgments. The authors thank Ambar Pal for his valuable comments and contribution to the development of this work. This research work was supported by DARPA GARD Program HR001119S0026-GARD-FP-052.

References

[1] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018. 3

[2] Qi-Zhi Cai, Min Du, Chang Liu, and Dawn Xiaodong Song. Curriculum adversarial training. *ArXiv*, abs/1805.04807, 2018. 6

[3] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017. 2

[4] João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the Kinetics-700 human action dataset. *ArXiv*, abs/1907.06987, 2019. 4

[5] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 1

[6] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv: Computer Vision and Pattern Recognition*, 2019. 6

[7] Kenneth T. Co, Luis Muñoz-González, and Emil C. Lupu. Procedural noise adversarial examples for black-box attacks on deep convolutional networks. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019. 2

[8] Guneet S. Dhillon, Kamyar Azizzadenesheli, Zachary Chase Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. *ArXiv*, abs/1803.01442, 2018. 3

[9] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. Spatiotemporal residual networks for video action recognition. In *NIPS*, 2016. 1

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015. 1, 2, 3

[11] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. *ArXiv*, abs/1711.00117, 2018. 3

[12] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 1, 4

[13] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[14] Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, and Jinwen He. Towards security threats of deep learning systems: A survey. *IEEE Transactions on Software Engineering*, 2021. 3

[15] Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvari. Learning with a strong adversary. *ArXiv*, abs/1511.03034, 2015. 2, 4

[16] Nathan Inkawhich, Matthew J. Inkawhich, Yiran Chen, and Hai Helen Li. Adversarial attacks for optical flow-based action recognition classifiers. *ArXiv*, abs/1811.11875, 2018. 1, 2

[17] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yugang Jiang. Black-box adversarial attacks on video recognition models. *Proceedings of the 27th ACM International Conference on Multimedia*, 2019. 1

[18] Guoqing Jin, Shiwei Shen, Dongming Zhang, Feng Dai, and Yongdong Zhang. Ape-gan: Adversarial perturbation elimination with gan. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3842–3846, 2019. 2, 3, 6

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. 3

[20] Hirokatsu Kataoka, Tenga Wakamiya, Kensho Hara, and Yutaka Satoh. Would mega-scale datasets further enhance spatiotemporal 3d cnns? *ArXiv*, abs/2004.04968, 2020. 1, 4

[21] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *ArXiv*, abs/1611.01236, 2017. 1, 3

[22] Wei-An Lin, Chun Pong Lau, Alexander Levine, Ramalingam Chellappa, and Soheil Feizi. Dual manifold adversarial robustness: Defense against lp and non-lp adversarial attacks. *ArXiv*, abs/2009.02470, 2020. 3

[23] Shao-Yuan Lo and Vishal M. Patel. Defending against multiple and unforeseen adversarial videos. *ArXiv*, abs/2009.05244, 2020. 1

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2018. 1, 2

[25] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018. 8

[26] Nicolas Papernot, Patrick Mcdaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597, 2016. 3

[27] Roi Pony, Itay Naeh, and Shie Mannor. Over-the-air adversarial flickering attacks against video recognition networks. In *CVPR*, 2021. 1, 2

[28] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, 2020. 6, 7

[29] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *ArXiv*, abs/1805.06605, 2018. 3

[30] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *arXiv: Computer Vision and Pattern Recognition*, 2019. 5

[31] Uri Shaham, Yutaro Yamada, and Sahand N. Negahban. Understanding adversarial training: Increasing local stability of

supervised models through robust optimization. *Neurocomputing*, 307:195–204, 2018. 3

[32] Yash Sharma and Pin-Yu Chen. Attacking the madry defense model with l1-based adversarial examples. *ArXiv*, abs/1710.10733, 2018. 3

[33] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *ArXiv*, abs/1710.10766, 2018. 3

[34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *ArXiv*, abs/1212.0402, 2012. 4

[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2014. 1, 2, 3

[36] Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. *ArXiv*, abs/1904.13000, 2019. 5, 6

[37] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick Mcdaniel. Ensemble adversarial training: Attacks and defenses. *ArXiv*, abs/1705.07204, 2018. 3

[38] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv: Machine Learning*, 2019. 6

[39] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *ArXiv*, abs/1601.06759, 2016. 3

[40] Xingxing Wei, Jun Zhu, and Hang Su. Sparse adversarial perturbations for videos. *ArXiv*, abs/1803.02536, 2019. 1, 2

[41] Chaowei Xiao, Ruizhi Deng, Bo Li, Taesung Lee, Ben Edwards, Jinfeng Yi, Dawn Xiaodong Song, Mingyan D. Liu, and Ian Molloy. Advit: Adversarial frames identifier based on temporal consistency in videos. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3967–3976, 2019. 1

[42] Hongyang R. Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019. 1

[43] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan S. Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *ICML*, 2020. 6

[44] Yuekai Zhang, Ziyan Jiang, Jesús Villalba, and Najim Dehak. Black-box attacks on spoofing countermeasures using transferability of adversarial examples. In *INTERSPEECH*, 2020. 2