AdvIT: Adversarial Frames Identifier Based on Temporal Consistency In Videos

Chaowei Xiao [1] * Ruizhi Deng [2] Bo Li [3] Taesung Lee [4]
Benjamin Edwards[4] Jinfeng Yi [5] Dawn Song [6] Mingyan Liu [1] Ian Molloy[4]
[1] University of Michigan, Ann Arbor [2] Simon Fraser University [3] UIUC
[4] IBM Research AI [5] JD.com [6] UC Berkeley

Abstract

Deep neural networks (DNNs) have been widely applied in various applications. However, DNNs are found to be vulnerable to adversarial examples. While several defense and detection approaches are proposed for static image classification, many security-critical tasks use videos as their input and require efficient processing. In this paper, we propose an efficient and effective method AdvIT to detect adversarial frames within videos against different types of attacks based on temporal consistency property of videos. In particular, we apply optical flow estimation to the target and previous frames to generate pseudo frames and evaluate the consistency of the learner output between these pseudo frames and target. High inconsistency indicates that the target frame is adversarial. We conduct extensive experiments on various learning tasks including video semantic segmentation, human pose estimation, object detection, and action recognition, and demonstrate that we can achieve above 95% adversarial frame detection rate. To consider adaptive attackers, we show that even if an adversary has access to the detector and performs a strong adaptive attack based on the state of the art expectation of transformation method, the detection rate stays almost the same. We also tested the transferability among different optical flow estimators and show that it is hard for attackers to attack one and transfer the perturbation to others. In addition, as efficiency is important in video analysis, we show that AdvIT can achieve real-time detection.

1. Introduction

Deep neural networks (DNNs) have been widely studied and have shown impressive performance in many tasks [15, 30, 31]. However, recent studies have shown that DNNs are vulnerable to *adversarial examples* [5, 6, 9, 17, 28, 35, 40–43] which are carefully crafted input instances targeted at leading machine learning models to produce attacker controlled errors in the output. This raises a number of security concerns in real-world machine learning based applications such as self-driving cars and surveillance [5, 6, 14, 28, 32].

While currently most research on adversarial examples focuses on static images, DNNs on videos is a particularly interesting and important domain, as attacks against many applications have the potential to cause serious physical and financial damage. For example, one application of DNNs to video is in autonomous vehicles; DNNs are used to identify other cars, road markings, street signs, and pedestrians. An adversarial attack forcing a network to classify a stop sign as a speed limit sign could easily cause a crash. Recently Wei et.al [38] proposed an adversarial attack targeting on action recognition task in videos which again emphasizes the vulnerabilities of learners for videos.

Several defense or detection methods have been proposed on static images but most of them are defeated by adaptive attacks [4, 8, 22, 23]. As a result, directly applying existing defenses on static images to videos is not robust nor efficient considering the high requirements for video processing. While general defense approach is hard, leveraging special properties of data source (e.g. videos) and enhancing model robustness is possible. In this paper, we propose *AdvIT* : the first adversarial frame identifier for videos based on temporal consistency. In particular, we allow the attacker to have white-box access to the target DNNs and add adversarial perturbation to one or more frames. Here we consider two types of attacks: *independent frame attack* which adds adversarial perturbation to selected frames independently (e.g. Houdini and DAG [10, 44]); and *temporal continuity attack* which generates perturbation considering the continuity among video frames (e.g. sparse and universal attack [38]). Our temporal consistency-based detection framework *AdvIT* is shown in Fig. 1. Given a target frame within a video, we first estimate the optical flows between the target $\boldsymbol{X}_t$ and its previous frames $(\boldsymbol{X}_{t-1}, \cdots, \boldsymbol{X}_{t-k})$. We then fuzz the estimated optical flows with small randomness $\alpha \sim \mathcal{N}(0, \sigma^2)$ and transform the previous frames as “pseudo frames” and check the consistency between the outputs of learning tasks based on the pseudo frames and the original target. We find that the frame transformation described above preserves the temporal consistency of learn-

*This work was performed when Chaowei Xiao was at IBM ing results if the target frame is benign, while weakens the pseudo frames' adversarial behaviour if the target frame is adversarial. This process is independent with the fact that whether the previous frames are adversarial or not. Therefore, we can leverage the prediction results of the pseudo frame as a reference and check its consistency to detect if the target frame is adversarial.

To demonstrate the effectiveness of *AdvIT* , we test our approach on four major video based tasks including semantics segmentation, human pose estimation, object detection and action recognition. Different state of the art attack approaches including Houdini, DAG, Sparse adversarial perturbation [10, 38, 44] are evaluated. We show that our approach can detect adversarial frames with above 95% detection rate. We also show that because of the large sample space of randomness added to the optical flow, even attacks that are aware of our detector would be unfeasible. As shown in the experimental section, our detection pipeline remains robust even under the proposed strategic adaptive attacks. We perform transferability analysis for different optical flow estimators and demonstrate that it is hard to transfer perturbation generated against one estimator to another. In addition, we analyze the performance of optical flow estimator and conclude that our approach does not require an accurate estimator to achieve high detection rate.

Our detection approach has several advantages compared to existing approaches: 1) we do not require time-consuming retraining of machine learning models as most of existing defenses [23, 36]; 2) our detector does not compromise the performance of the learning tasks; 3) due to the randomness injected, it is hard to perform adaptive attacks against the detection method; 4) we do not require the optical flow estimation to be differentiable, which ensures its wide application;

Contributions (1) We propose to leverage the temporal consistency in videos and and randomness to develop an efficient and effective approach *AdvIT* that detects adversarial frames with above 95% detection rate. To the best of our knowledge, this is the first work to apply optical flow to quantitatively estimate the consistency of learning tasks on videos and use it to detect adversarial behaviours. (2) We conduct extensive experiments and analyses to identify adversarial frames within videos against different state-of-the-art attacks on video learning tasks including semantic segmentation, human pose estimation, object detection, and action recognition. We show that *AdvIT* outperform potential baselines significantly. (3) We propose strong adaptive attacks against our detection method and show that it is robust against the proposed attacks which assume adversaries are aware of the detection mechanism. 4) We evaluate the transferability among different optical flow estimators and show that adversarial attacks rarely transfer among them, which motivate us to embed a non-differential optical flow estimator into our detection system.

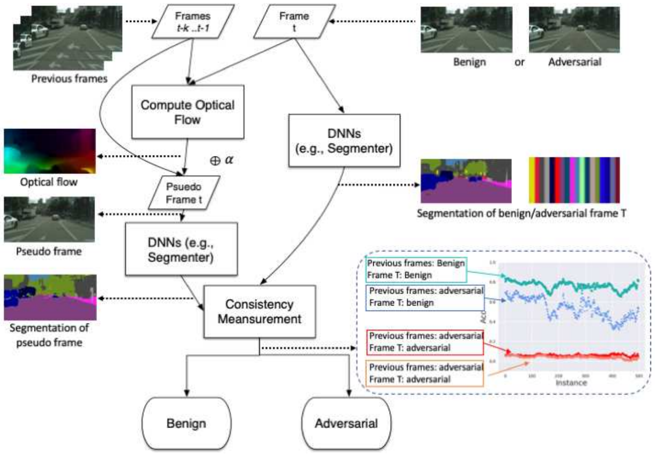


Figure 1: Pipeline of the proposed temporal consistency based adversarial frame identifier: *AdvIT* .

2. Related work

In this section we will provide brief introduction of current adversarial attacks on videos, as well as potential defense approaches against adversarial examples.

Learning for videos. Deep neural networks have been successfully applied to video in a number of supervised computer vision tasks including: *Semantic segmentation, object detection and human pose estimation.* Fully convolution networks [21] propose an end-to-end model that first down-samples the feature map and then up-sample to generate a pixel-wise class score map for semantic segmentation. [45] improves this pipeline by introducing dilated convolution that increases the receptive field size without decreasing its resolution. Object detection has been accomplished using R-CNN models that adopt a proposal and prediction pipeline [16, 31] for object detection and semantic segmentation. YOLO [29, 30] models only make predictions on a fixed set of bounding boxes restricted by the grid of feature map and a pre-defined set of anchors, but this limitation allows it to achieves real-time performance. Stacked Hourglass Networks [25] achieve state of the art performance on the task of single person humans pose estimation through using a repeated top-down and bottom-up model and capturing information at all scales.

Adversarial attacks. Adversarial examples have been heavily explored in classification task [5, 6, 9, 17, 28, 35, 40, 42, 43]. DAG [44] and Houdini [10] have both generated imperceptible changes to inputs to create attacker controlled outputs against segmentation and object detection tasks. DAG [44] proposes an iterative gradient based attack methods to attack all pixels until most of the pixels have been identified as target classes for semantic segmentation while it attacks all proposed bounding boxes until they are misclassified as target classes. Houdini [10] proposes a optimization based attack algorithm by introducing a surrogate loss function. Adversarial attacks have recently been extended into the domain of video data. Sparse adversarial perturbation [38] demonstrates a method to generate universal adversarial perturbations against action recognition model for videos. The method also achieves temporal sparsity by imposing a temporal mask upon the perturbation.

Defenses against adversarial examples. Various detection and defense methods have also been explored against adversarial examples in image classification, though they have not considered or tested video inputs. Adversarial training [17] and its variations [23, 36] have generally been more successful, but usually come at the cost of accuracy and increased training time [37]. Recently, Athalye et al. [4] successfully generated adversarial examples in the presence of detection and defense strategies. Xiao et al. [39] proposed a method to detect adversarial examples on semantic segmentation using spatial information. However, these methods all focus on static image behaviour and is not directly applicable to object detection and human pose estimation. Currently no defense or detection methods have been studied that can be applied across video-based tasks, such as human pose estimation and object detection task.

3. Adversarial frame identifier via temporal consistency: *AdvIT*

In this section, we first formally define the problem. Then, we provide an overview of the proposed approach advIT, and discuss each step in detail.

Formally, we define the problem as follows: Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t$ be the sequence of image frames of a (streaming) video, and $\boldsymbol{X}_t$ is the target frame. Let g be a learner with output $g(\boldsymbol{X}_t) = \boldsymbol{Y}_t$. In this context, the attacker can inject small perturbation (*i.e.*, $\boldsymbol{X}_i \leftarrow \boldsymbol{X}_i + \boldsymbol{\epsilon}_i$) to one or more frames to achieve $g(\boldsymbol{X}_t) = \boldsymbol{Y}^*$ where $\boldsymbol{Y}^*$ is the adversarial target depending on the learning task. Our goal is to determine whether the target frame $\boldsymbol{X}_t$ is adversarial without any other knowledge except for the previous k frames $\boldsymbol{X}_{t-k}, \ldots, \boldsymbol{X}_{t-1}$, and it is not clear whether the previous k frame are benign or adversarial.

Threat Model In this work we mainly focus on two types of frame based adversarial attacks: *independent frame attack* and *temporal continuity attack*, both of which aim to add perturbation to one or more frames and therefore mislead the target learner. We believe such frame based attack could lead to severe consequences given the fact that frame based approaches are the most effective and commonly used ones in videos [7, 30]. In particular, *independent frame attack* includes Houdini [10] and DAG [44] attacks for video segmentation, object detection and human pose estimation tasks; and *temporal continuity attack* includes Sparse attack [38] on action recognition and Universal perturbation [24] on the previous three tasks. To avoid trivial detection, an adversary needs to constraint the magnitude of perturbation. Without loss of generality, we use l_2 to bound the added perturbation. Some video attack examples for the considered four learning tasks are shown in Fig 2 and Fig 3.

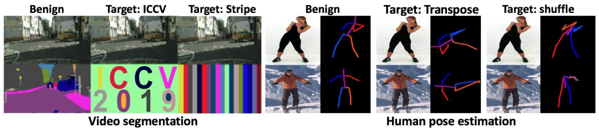


Figure 2: Benign and adversarial frames generated on Cityscapes and Davis Challenge 17 dataset for video segmentation and human pose estimation respectively.

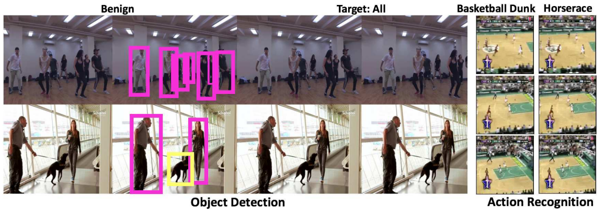


Figure 3: Benign and adversarial frames generated on MPII and UCF-101 for video object detection and action recognition respectively.

Overview of Method Our detection algorithm is based on temporal consistency of videos. Since the next frame in the video is the continuation of the objects and the scene after their small movements, they can be mostly reconstructed from the current frame if we can compute such movements. Due to this continuity, we should have consistent learning outputs from neighbor frames [19]. However, this may not be true when the next frame contains adversarial perturbation which may interrupt such property.

We call the reconstructed frame as a *"pseudo frame"*, and can *validate* the learning output temporal consistency by comparing the output of the *to-be-verified* target frame and the pseudo frame. Specifically, we observe the following properties regarding the validation of the temporal consistency with the pseudo frame. First, since adversarial perturbation is very specific to the frame, a newly reconstructed pseudo frame is much less affected by the adversarial perturbation. Its output is very close to that of a benign frame. Second, adversarial perturbation in the target frame breaks the output temporal continuity compared to the output of the pseudo frame. Thus, if the target frame is adversarial, we can observe that the temporal consistency of the output does not hold.

Based on this observation, we propose the adversarial frame detection framework that tests temporal consistency of the outputs as shown in Fig 1. First, we generate pseudo frames based on each of the k previous frames ($\boldsymbol{X}_{t-k} \cdots \boldsymbol{X}_{t-1}$) by estimating optical flow (O_F) from each to the target frame X_t, and adding certain random transformation $\alpha \sim \mathcal{N}(0, \sigma^2)$. Then, we run the learner (*e.g.*, segmenter) on the pseudo frames and the target frame to get prediction results. Finally, we compare their results to test if the temporal consistency is satisfied. Note that our method is independent of the adversarial behaviour of the previous frames: they can either be adversarial or benign. The bottom-right box in Fig. 1 shows the four scenarios. Next we will introduce the two components of the proposed method in details: (1) Pseudo frame generation; (2) temporal consistency based test.

3.1. Pseudo Frame Generation

We combine two types of transformations to generate the pseudo frames: optical flow and random transformation. First, optical flow is a transformation caused by movements of the viewer or the objects in the scene. This technique is able to reconstruct subsequent frames and is the basis of a number of video codecs, such as MPEG-2 [27]. Machine learning models have also successfully been trained to generate the vector field $\boldsymbol{V}$ when trained on videos [12, 19]. In particular, optical flow has been applied to estimate the instantaneous 2D velocity of visible surface points from time-varying image signal. Traditional optical flow estimation methods usually involve solving optimization problems based on assumptions of consistency, smoothness of intensity, and gradients [20].

An optical flow estimator is a function F which generates a vector field V indicating the direction and distance to move pixels within an image. Deep-learning-based optical flow estimation models have been widely studied on different video tasks. Dosovitskiy et.al [12] firstly applied deep neural network to estimate the optical flow. Here we leverage the DNNs based optical flow [19] to characterize the continuity among video frames. Given two temporally close frames, we can quantify the motion using an optical flow estimation algorithm. By applying the optical flow to the first input frame, we can reconstruct the second one.

Formally, let $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$ be a pair of temporally close frames in a video. An optical flow between the two frames is a vector field $O_F = (\boldsymbol{\Delta u}, \boldsymbol{\Delta v})$ that describes the displacement of pixels between the frames and we denote an image generated by applying flow as $\hat{\boldsymbol{X}}_{s\to t}$. The goal of an optical flow algorithm is to minimize the error of the generated image $\hat{\boldsymbol{X}}_{s\to t}$ and the actual frame $\boldsymbol{X}_t$. We obtain $\hat{\boldsymbol{X}}_{s\to t}$ by sampling pixel intensities from $\boldsymbol{X}_s$; the pixel in $\hat{\boldsymbol{X}}_{s\to t}$ at location (i, j) corresponds to the pixel at location $(u, v) = (i + \boldsymbol{\Delta u}(i, j), j + \boldsymbol{\Delta v}(i, j))$ in image $\boldsymbol{X}_s$. As $(\boldsymbol{\Delta u}, \boldsymbol{\Delta v})$ can be fractional numbers and (u, v) does not necessarily lie on the integer coordinate grid, the pixel intensity can be sampled via bilinear sampling.

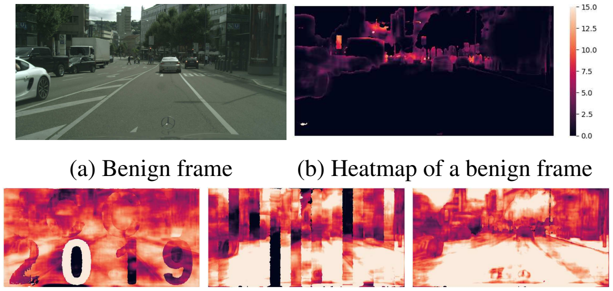
$$\hat{\boldsymbol{X}}_{s\to t}(i, j) = \sum_{(i', j') \in N(u,v)} \boldsymbol{X}_s(i', j')(1 - |u - i'|)(1 - |v - j'|) \quad (1)$$

where $N(u, v)$ stands for the indices of the 4-pixel neighbors at location (u, v) (top-left, top-right, bottom-left, bottom-right). $\boldsymbol{X}.(i, j)$ represents the pixel value at location (i, j).

To further combat the creation of adversarial perturbation, we add randomness $\boldsymbol{\alpha} \sim \mathcal{N}(0, \sigma^2)$ to the flow field $(\boldsymbol{\Delta u}, \boldsymbol{\Delta v})$ to generate the pseudo frames. This randomness can also help make adaptive attacks harder, where an adversary has full knowledge about the detection.

Algorithm 1: Temporal Consistency Based Test

```
input:  target frame in a video$\boldsymbol{X}_t$;
        previous K frames of $\boldsymbol{X}_t$: $\boldsymbol{X}_{t-k}, \ldots, \boldsymbol{X}_{t-1}$;
        optical flow estimation model flow;
        machine learning model $g$;
        consistency evaluation function $f$;
output: Continuity metric $c$;
Initialization : cs ←[],
  $w \leftarrow x.width, h \leftarrow x.height, \boldsymbol{Y}_t \leftarrow g(\boldsymbol{X}_t)$;
for $s \leftarrow t-1$ to $t-k$ do
    $(\boldsymbol{\Delta u}, \boldsymbol{\Delta v}) \leftarrow \mathbf{flow}(\boldsymbol{X}_s, \boldsymbol{X}_t)$;
    /* add randomness to optimal flow     */;
    $(\boldsymbol{\Delta \tilde{u}}, \boldsymbol{\Delta \tilde{v}}) \leftarrow (\boldsymbol{\Delta u}, \boldsymbol{\Delta v}) + \boldsymbol{\alpha}$;
    /* generate pseudo frame $\boldsymbol{X}_{T-k}$     */;
    $\hat{\boldsymbol{X}}_{s\to t} \leftarrow \mathbf{warp}((\boldsymbol{\Delta \tilde{u}}, \boldsymbol{\Delta \tilde{v}}), \boldsymbol{X}_s)$;
    $\hat{\boldsymbol{Y}}_{s\to t} \leftarrow g(\hat{\boldsymbol{X}}_{s\to t})$;
    /* measure consistency information     */;
    $\mathbf{cs} \overset{+}{\leftarrow} f(\hat{\boldsymbol{Y}}_{s\to t}, \boldsymbol{Y}_t)$;
end
$c \leftarrow \mathbf{Mean}(\mathbf{cs})$;
Return: c
```



(a) Benign frame (b) Heatmap of a benign frame

(c) ICCV (d) Stripe (e) Remapping

Figure 4: Heatmap of per-pixel cross-entropy. (a) and (b) show a benign frame and the corresponding per-pixel cross entropy between the prediction of its pseudo frames and itself. The rest show similar per-pixel cross entropy for adversarial frames with different targets. The labels indicate their adversarial targets.

3.2. Temporal Consistency Based Test

To quantitatively demonstrate the difference of temporal consistency between adversarial and benign cases, we

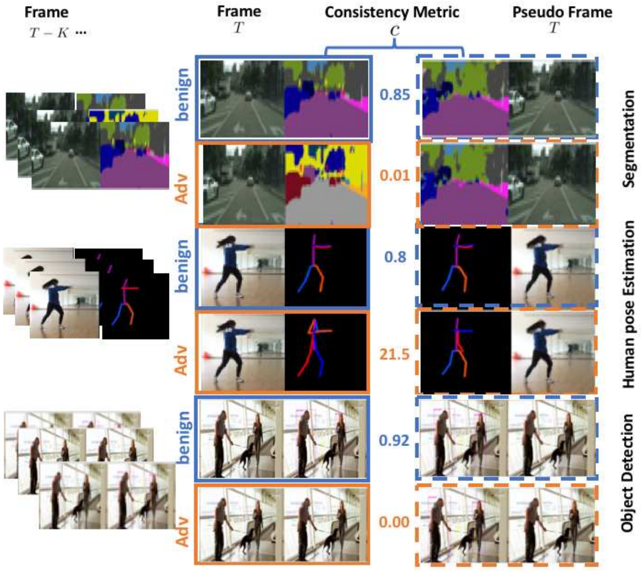


Figure 5: Examples of consistency measurement based on *AdvIT* for various video tasks. The first column indicates previous Frames. The second column indicates the current frame and corresponding prediction result. The last column indicates a sampled pseudo frame and corresponding prediction. The consistency metric C shows quantitative results for different learning tasks. Note that higher C for segmentation and object detection means higher consistency, while lower C indicates more consistent for Human pose estimation since it is based on L_2 distance.

first use semantic segmentation as an example here to illustrate our findings. Suppose given $k+1$ consecutive frames from video, $\boldsymbol{X}_{t-k}, \ldots, \boldsymbol{X}_{t-1}, \boldsymbol{X}_t$, we estimate the O_F between each of the previous k frames $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$, where $s = t-k, ..., t-1$. Next, we add randomness α to each O_F to reconstruct pseudo frame $\hat{\boldsymbol{X}}_{s\to t}$ from $\boldsymbol{X}_s$. These pseudo frames and $\boldsymbol{X}_t$ are then sent as input to learning model g. We normalize the output of g by the softmax function so that the prediction of every pixel is a vector indicating the probability of the pixel belonging to every class. The prediction results of the learning models are denoted by $\hat{\boldsymbol{Y}}^{\boldsymbol{d}}_{s\to t}$ and $\boldsymbol{Y}^{\boldsymbol{d}}_t$ respectively. We compute the average for the prediction vectors of k pseudo frames, $\boldsymbol{S} = \frac{1}{k}\sum_{s=t-k}^{t-1} \hat{\boldsymbol{Y}}^{\boldsymbol{d}}_{s\to t}$, and we use $\boldsymbol{S}(i,j)[m]$ to indicate the averaged probability of pixel (i,j) being predicted to be class m in pseudo frames. Based on the k pseudo frames, we calculate the cross entropy $\boldsymbol{E}$ between $\boldsymbol{S}$ and $\boldsymbol{Y}_t$ as

$$\boldsymbol{E} = \sum_m -\boldsymbol{Y}_t^d[m] \circ \log \boldsymbol{S}[m]$$

where $\circ$ denotes Hadamard product. We visualize the cross entropy $\boldsymbol{E}$ in Fig. 4: Fig. 4c to Fig. 4e show the heatmaps of cross entropy when $\boldsymbol{X_t}$ is adversarial examples with different attack targets, while Fig. 4b shows the heatmap of cross entropy when $\boldsymbol{X_t}$ is benign. It is clear that for the benign instance, the prediction results for most pixels are consistent except for small regions around boundaries of the objects; while for the adversarial targets, most of the pixels show inconsistent prediction results. We also observe that whether $\boldsymbol{X}_{t-k}, \ldots, \boldsymbol{X}_{t-1}$ are benign or adversarial has little impact on the prediction consistency for $\boldsymbol{X_t}$.

Based on such observation, we proposed leveraging the temporal consistency information to distinguish adversarial frames in videos and provide the following detailed algorithm. Without loss of generality, we assume $\boldsymbol{X}_t$ is a current frame and our goal is to detect whether the current frame $\boldsymbol{X}_t$ is adversarial. Given an optical flow model, we denote the optical flow between previous frame $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$ by $O_F = (\boldsymbol{\Delta u}, \boldsymbol{\Delta v})_{s\to t}$. Randomness $\boldsymbol{\alpha}$ is added to the optical flow $(\boldsymbol{\Delta u}, \boldsymbol{\Delta v})_{s\to t}$ to get new optical flow $(\boldsymbol{\Delta\tilde{u}}, \boldsymbol{\Delta\tilde{v}})_{s\to t}$ where $(\boldsymbol{\Delta\tilde{u}}, \boldsymbol{\Delta\tilde{v}})_{s\to t} = (\boldsymbol{u}, \boldsymbol{v})_{s\to t} + \boldsymbol{\alpha_{u,v}}$. After obtaining $(\boldsymbol{\Delta\tilde{u}}, \boldsymbol{\Delta\tilde{v}})_{s\to t}$ for $s = t-k, ..., t-1$, we generate the pseudo frames $\hat{\boldsymbol{X}}_{s\to t}$. We then calculate the consistency metric c between $\boldsymbol{Y}_t = g(\boldsymbol{X}_t)$ and $\boldsymbol{Y}_{s\to t} = g(\hat{\boldsymbol{X}}_{s\to t})$ respectively with a scalar consistency function f to determine whether $\boldsymbol{X}_t$ is an adversarial frame or not, where $c = \frac{1}{k}\sum_{s=t-k}^{t-1} f(\boldsymbol{Y}_{s\to t}, \boldsymbol{Y}_t)$. The algorithm of this temporal continuity based method is shown in Algorithm 1, where **warp** is achieved via bilinear sampling as defined in Eq. 1.

We adopt different consistency measurement function f for various learning tasks: (1) Segmentation: Pixel-wise accuracy[1]. (2) Human Pose Estimation: Average L_2 distance over all key joints. Note that, high distance indicates low consistency. (3) Object Detection: mIoU between bounding boxes of pseudo frames and the current frame. (4) Action Recognition: the average of forward and backward KL divergence between the two categorical distributions. The detailed algorithm to calculate the mIoU is shown in supplementary.

Fig. 5 shows the examples of our detection method. We can observe that the inconsistency between the target frame and corresponding pseudo frames is high for an adversarial frame and low for benign, and this conclusion holds for various learning tasks on video.

4. Experimental Results

In this section, we present experimental results on detecting adversarial frames with *AdvIT* against different attacks for four learning tasks on videos, including video semantic segmentation, human pose estimation, object detection, and action recognition. We show that our adversarial frame detection method is robust even under a strong adaptive attack

[1] Pixel-wise accuracy and mIoU provide similar results and the former is much more computationally efficient.

where the adversary has perfect knowledge of the learning model and detection mechanism.

4.1. Implementation Details

Semantic segmentation For semantic segmentation task, we use CityScapes dataset which consists of high-resolution (1024x2048) outdoor videos captured from a moving car. Attacks against the CityScapes dataset pose a realistic threat to recognition models in real-world applications, in particular autonomous driving. We adopt the state-of-the-art Dilated Residual Network [45] model with DRN-D-22 architecture which is trained on the CityScapes dataset. The mean Intersection Over Union (mIoU) of this model on pristine data is 66.7. We demonstrate our results on video clips from the same dataset, each consisting of 100 high-resolution frames shot at a frame rate of 17Hz. We evaluate over three different adversarial targets: “Remapping”, “Stripe”, and “ICCV 2019”. The details of these targets are shown in supplementary materials. We generate adversarial frames based on two state-of-the-art attack methods: Houdini [10] and DAG [44]. Each attack was run with a maximum perturbation of $l_2 = 0.03$ with input frames scaled from $[0, 1]$ until 98% pixel-wise accuracy of the target was achieved. We select $\sigma = 0.002$ for detection.

Human Pose Estimation In human pose estimation task, we attack the Stacked Hourglass Network model [25] trained on MPII human pose dataset [3]. The two-stack model we use is pretrained on the MPII dataset [3] and achieved a mean PCHk score of 86.95 on the MPII validation dataset at the threshold of 0.5. We attacked 3 clips of video data from the MPII human pose estimation dataset and YouTube [2] using the Houdini algorithm which is the current the-state-of-art. The attack targets we choose are “Transpose” and “Shuffle”. “Transpose” means transposing coordinates of benign image predictions. “Shuffle” means shuffling the joints of the pose predictions. We select $\sigma = 0.02$ during detection.

Object Detection We use YOLOv3 [30] as our target model for the object detection task. We select two video clips randomly from DAVIS Challenge 17 dataset [1] to perform attacks on. We select two targets for simplicity: “All” and “Person”. “All” means removing all of the bounding boxes in the images while “Person” means removing only the bounding boxes of persons in images. Such attacks potentially show that every image taken in the real surveillance system can be attacked to the scene without any object or without any person which brings severe security concerns for current surveillance systems based on object detection algorithms. We use DAG algorithm which is the current the-state-of-the-art algorithm, to attack the YOLOv3. We run the attack with a maximum perturbation of 0.03 with input frames scaled from $[0, 1]$ until we achieve 100% removal of the target objects. We select $\sigma = 0.002$ during detection.

Action Recognition This task and the corresponding attack method Sparse takes the video temporal continuity into account. The target model makes action predictions based on a whole clip of a video instead of processing individual frame, i.e.

$$Y = F(X_1, X_2, \dots, X_N)$$

We use the CNN+RNN model used in [38] as video action recognition model and also apply the state of the art attack Sparse to generate adversarial video clips. The model is trained on the UCF-101 dataset [33] to predict the action from 101 classes, and uses a Inception V2 model [34] to extract features from each frame. Given a recognition model $\mathbf{F}_\theta$, several of video clips $\{\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_N\}$ and their corresponding adversarial target labels $\{y_1, y_2, \dots, y_N\}$, the attack optimizes the following objective:

$$\underset{\mathbf{E}}{\operatorname{argmin}}\, \lambda ||\mathbf{M} \cdot \mathbf{E}||_p + \frac{1}{N} \sum_{i=1}^{N} l(\mathbf{1}_{y_i^*}, \mathbf{F}_\theta(\mathbf{C}_i + \mathbf{M} \cdot \mathbf{E}))$$

where $\mathbf{E}$ is a universal adversarial perturbation and $\mathbf{M}$ is a predefined temporal mask to enforce sparsity. $l(\cdot)$ represents the classification loss. The attack aims to generate a universal perturbation for all N clips, and in our experiment the recognition model mis-classified 13245 videos out of 13320 samples. We use *AdvIT* to detect adversarial video clips instead of adversarial frames for this task.

Task	Attack Method	Target	Defense Method	Detection (k)		
				1	3	5
Semantic Segmentation	Houdini	Stripe	*Replacement* *JPEG* *AdvIT*	50% **100%** **100%**	50% - **100%**	50% - **100%**
Human Pose Estimation	Houdini	Shuffle	*Replacement* *JPEG* *AdvIT*	50% 98% **100%**	50% - **100%**	50% - **100%**
Object Detection	DAG	Person	*Replacement* *JPEG* *AdvIT*	50% 60% **98%**	50% - **99%**	50% - **100%**

Table 1: Comparison of detection results (AUC) against different attacks for *AdvIT* and baseline methods.

4.2. Temporal Consistency Based Detection

Here we evaluate the detection performance of *AdvIT* on different video tasks comparing with other baseline methods. Following Algorithm 1, given a frame, we first generate pseudo frames for its previous k frames and then calculate the consistency metric between the frame and these pseudo frames. Based on the distribution of consistency metric C, we identify the adversarial frames if C is low, and vice versa.

We evaluate *AdvIT* against two types of frame based adversarial attacks: *independent frame attack* and *temporal continuity attack* respectively. For both scenarios, we report the Area Under Curve (AUC) of Receiver Operation Characteristic Curve (ROC) of *AdvIT* and baselines.

Detecting independent frame attack *Independent frame attack* includes Houdini [10] and DAG [44] on three video tasks: semantic segmentation, human pose estimate and object detection. Since each frame is independent for attackers, the attacker can decide whether to attack the previous frame so the the previous frames can be either adversarial or benign. Our method aims to identify whether current frame is adversarial, regardless the status of previous frames. Therefore, we test our method and report the results under various conditions: previous frames are purely benign, adversarial, or mixture. The result is shown in Tab. 1 and other result is shown in supplementary materials. Note that as this is the first work to detect adversarial frames within a video, there is no existing detection methods dedicated to video to compare with. To demonstrate the effectiveness of *AdvIT* , we, instead, compared our method with two baselines: a traditional static image based method, JPEG compression [11, 13, 18] shown as *JPEG* and *Replacement* in Tab. 1. *JPEG* compresses the current frame to generate a "pseudo frame" and then calculates the consistency metric between the current and "pseudo frame". *Replacement* directly leverages the temporal continuity by replacing the current frame with previous frames to generate "pseudo frames". It then calculates the continuity metrics between current frame and the "pseudo frame". We evaluate the performance by aggregating previous k frames, where $k \in \{1, 3, 5\}$. Note that *JPEG* only considers current frame so it only applies for $k = 1$.

From Tab. 1. We also observed that even *JPEG* achieves high detection rate on Semantic segmentation and Human pose estimation, it is less robust than *AdvIT* and performs worse on object detection. It indicated that the perturbation on object detection might be subtle against compression. Compared with the baseline methods, *AdvIT* shows promising detection performance (almost 100%) against independent frame attacks on various learning tasks under different scenarios. (More detection results against different attack targets are omitted in supplementary.)

To further illustrate the effectiveness of *AdvIT* , we show *AdvIT* on the two extreme cases: previous frames are adversarial or benign. We observe that *AdvIT* achieve almost 100% success rate for identifying adversarial frames all kinds of settings without requiring knowledge about whether previous frames are adversarial. The complete results are deferred to the supplementary.

Detecting temporal continuity attack Considering attacks that also take temporal continuity of videos into account, we evaluate the effectiveness of *AdvIT* on *temporal continuity attack*, Sparse attack [38] on video action recognition and universal perturbation for the previous three tasks. For universal perturbation, we generate universal perturbation for 5 frames within videos. We extend the DAG and Houdini methods to generate universal perturbation. Tab. 2 shows the detection results of *AdvIT* against such attacks. We observe that *AdvIT* can defend against the universal perturbation with 100% detection rate on different video tasks, which implies that universal perturbation can not transfer to pseudo frames. Though Sparse attack [38] aims to generate sparse and continuous perturbation for videos, *AdvIT* can still achieve high detection rate by leveraging both temporal consistency and randomness. Note that the detection rate increases slightly with the growth of k, but there is no need to carefully tune k since $k = 1$ already achieves detection rate above 95%. In addition, to analyze the effectiveness of *AdvIT* against the adversarial attack with different strength, we conduct experiment by limiting the perturbation magnitude to 2, 16, 32 pixels (in range of [0,255]). The detailed results are shown in supplementary materials. It shows that the detection rate will decrease a bit with the magnitude increasing. But with ensemble of previous k frames, it is still effective.

Task	Attack Method	Target	Detection (k)		
			1	3	5
Semantic Segmentation	Universal	Strip	100%	100%	100%
Human Pose Estimation		shuffle	100%	100%	100%
Object Detection		all	100%	100%	100%
Action Recognition	Sparse	-	95%	96%	97%

Table 2: Detection results (AUC) against *temporal continuity attack*

4.3. Analysis of Adaptive Attacks

Given a detection method, it is important to evaluate it against a strong adaptive attacker who is aware of the detection mechanism. Thus, we conduct experiments to simulate the strong adaptive attacker we can think of to evaluate the robustness of *AdvIT* , assuming the attacker has complete access to our fully differentiable models. First, the attacker generates a perturbation that considers both the current and generated pseudo frames. Implementation details of adaptive attack will be included in supplementary material. Such attack will fail since during detection as we add randomness α to make the optical flow estimation harder. Thus, we allow the attacker to use the state of the art adaptive attack estimation method *expectation of transformation* to approximate potential randomness [4]. We follow the setting in [4], and randomly select 30 possible α in each iteration to optimize the perturbation. We select $l_2 = 0.03$ as the upper bound of the adversarial perturbation (pixel values are in range [0,1]), as perturbation larger than that would produce noticeable visual changes to human. The detection results against such adaptive attacks among different video tasks are shown in Tab. 3 as "Detection Adap". We observe that *AdvIT* can still achieve above 95% detection rate. We hypothesize that it is because (a) the high dimension of spatial randomness introduces large search space; (b) indirect changes to the pseudo frames during attack are not sufficient to manipulate the prediction of both the pseudo frames and target one.

Task	Target	Previous Frames	Detection Adap (k)			Detection Trans (k) (non-differential flow)		
			1	3	5	1	3	5
Semantic Segmentation	ICCV	Benign	100%	100%	100%	100%	100%	100%
		Adversarial	95%	97%	100%	100%	100%	100%
	Remapping	Benign	100%	100%	100 %	100%	100%	100%
		Adversarial	96%	96%	98%	100 %	100%	100%
Human pose estimation	Shuffle	Benign	96%	97%	97%	100%	100%	100%
		Adversarial	94%	97%	100%	100%	100%	100%
	Transpose	Benign	98%	99%	100 %	100%	100%	100%
		Adversarial	95%	95%	100 %	100 %	100%	100%
object detection	All	Benign	99%	100%	100%	100%	100%	100%
		Adversarial	99%	100%	100%	100%	100%	100%
	Person	Benign	98%	99%	100 %	100%	100%	100%
		Adversarial	95%	96%	97%	100 %	100%	100%

Table 3: Detection results (AUC) of adaptive attacks and transferability analysis.

Transferbility analysis In addition to the randomness for optical flow estimation process, in this section we try to analyze the transferability of the perturbation between different flow estimator. For instance, we allow the defender to use a non-differentiable flow estimator, while the attacker uses a differentiable one for the sake of attack convenience.

We substitute a non-differential flow estimator proposed by [20] in Algorithm 1 and evaluate its performance against an adversary who can approximate flow using a differential flow estimator FlowNet [12]. Such transferability based detection results are shown in Tab. 3 “Detection Trans (non-differentiable flow)". We can see that the transferability based adversarial perturbation generated for differential flow estimator does not transfer to non-differential flow estimator and the detection results are 100% across all k.

Optical Flow Estimator Next we will evaluate the impact of optical flow estimator on *AdvIT* . We calculate the accuracy of the applied flow estimator for different learning tasks, and the accuracy is around 1% in different scenarios after adding randomness α (detailed results are omitted to supplementary). This observation indicates that the high detection performance of *AdvIT* does not rely on very accurate optical flow estimator, which makes the proposed detection widely applicable.

Run-time Analysis For video based tasks, the frame process efficiency is important. Here we show that our adversarial detection approach *AdvIT* processes the videos with minimal overhead, compared with the source frame rate. Theoretically, our approach runs the model inference k times more; while running the model is the major cost, and k is a small constant, we have the same time complexity as the original learner's inference. Empirically, we measure the additional running time for different tasks using an Nvidia 1080Ti GPU. We present the run-time results of model inference, detection, and overhead (subtraction of the two) in Tab. 4. For human pose estimation and object detection, *AdvIT* has low overhead, 0.03 and 0.05 seconds respectively, yielding less than two frames of delay. The overhead for segmentation is higher at 0.4s, while the cost to run segmentation on the original input images is much larger, 2.58s on average, and our overhead is only 15.5%. When run in parallel with other real-time action recognition models [26, 30] and flow estimators, our detection pipeline can also achieve close to real-time performance.

Task	Inference	Detection	Overhead
Segmentation	2.58 ± 0.29	2.98 ± 0.27	0.4
Human Pose Estimation	0.02 ± 0.01	0.05 ± 0.01	0.03
Object Detection	0.04 ± 0.01	0.09 ± 0.01	0.05
Action Recognition	0.50 ± 0.01	0.52 ± 0.01	0.02

Table 4: Detection overhead of *AdvIT* (in seconds).

5. Discussion and Conclusion

We have present an effective and efficient adversarial frame detection method *AdvIT* for various video based tasks. We have used the state of the art learners, and challenged our detector with the best-known attacks. Further, we have done our best to identify scenarios where an adversary is aware of the detector and may seek to use information about the detector to circumvent it. Even with this strong adaptive adversary we are able to detect nearly all adversarial frames (above 95%). More sophisticated future attacks, which rely on different assumptions than those laid out here may be able to create adversarial frames while fooling our detector, which will be interesting future directions. Given the sequential nature of video, this work indicates that developing new attacks is likely to be more difficult in the video domain than in static image analysis.

Our experiments rely on video tasks where video is taken from continuous sequence. An video that contains “jump cuts" would likely introduce a small number of false positive frames into our detector, as these cuts would be unpredictable by the optical flow algorithm. However, continuous video is consistent with our applications (particularly autonomous driving and surveillance systems), and is unlikely to contain such cuts. It is also possible that we could detect such cuts, as the difference between the pseudo frames and the current frames would still be large, which needs further studies.

It should be noted that the goal of our work is to detects adversarial frames, but is not aimed at remediating or repairing them. Future work could include using pseudo frames as surrogates for suspicious frames or using pseudo frames along with detected adversarial frames to reform attacks.

Acknowledgement This work is partially supported by the Berkeley Deep Drive,the Center for Long-Term Cybersecurity and National Science Foundation under Grant CNS-1422211, CNS-1616575, IIS-1617767 and DARPA under Grant 00009970.

References

[1] Davis challenge 2017. `https://davischallenge.org/challenge2017.html`.

[2] Youtube video. `https://www.youtube.com/watch?v=vgusHl1Oue0`.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

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

[5] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. *CCS*, 2019.

[6] Yulong Cao, Chaowei Xiao, Dawei Yang, Jing Fang, Ruigang Yang, Mingyan Liu, and Bo Li. Adversarial objects against lidar-based autonomous driving systems. *arXiv preprint arXiv:1907.05418*, 2019.

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[8] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017.

[9] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017.

[10] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *Advances in Neural Information Processing Systems 30*, 2017.

[11] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.

[12] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE ICCV*, pages 2758–2766, 2015.

[13] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

[14] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE CVPR*, pages 580–587, 2014.

[17] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

[18] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. *ICLR*, 2018.

[19] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE conference on CVPR*, volume 2, page 6, 2017.

[20] Ce Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009.

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[22] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Michael E Houle, Grant Schoenebeck, Dawn Song, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[24] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1765–1773, 2017.

[25] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[26] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.

[27] Atul Puri, Richard Kollarits, and Barry Haskell. Basics of stereoscopic video, new compression results with MPEG-2 and a proposal for MPEG-4. *Sig. Proc.: Image Comm.*, 10(1-3):201–234, 1997.

[28] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditional image editing. *arXiv preprint arXiv:1906.07927*, 2019.

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[30] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[32] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernan-

des, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018.

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, 2012.

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[36] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *ICLR*, 2018.

[37] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy). *ICLR 2019*, 2018.

[38] Xingxing Wei, Jun Zhu, and Hang Su. Sparse adversarial perturbations for videos. *AAAI 2019*, 2018.

[39] Chaowei Xiao, Ruizhi Deng, Bo Li, Fisher Yu, Dawn Song, et al. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In *Proceedings of the (ECCV)*, pages 217–234, 2018.

[40] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *IJCAI*, 2018.

[41] Chaowei Xiao, Xinlei Pan, Warren He, Jian Peng, Mingjie Sun, Jinfeng Yi, Bo Li, and Dawn Song. Characterizing attacks on deep reinforcement learning. *arXiv preprint arXiv:1907.09470*, 2019.

[42] Chaowei Xiao, Dawei Yang, Bo Li, Jia Deng, and Mingyan Liu. Meshadv: Adversarial meshes for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6898–6907, 2019.

[43] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018.

[44] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *International Conference on Computer Vision*. IEEE, 2017.

[45] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.