AdvIT: Adversarial Frames Identifier Based on Temporal Consistency In Videos (Supplementary Material)

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

A. Implementation Details

Attack setting As described in the threat model section, we limit the $\mathcal{L}_2$ norm of the adversarial perturbation to be below 8 pixel in all settings because otherwise the the adversarial perturbation would be perceptible and can be easily detected. We set a max iteration of 1000 as the breakout condition.

Adversarial targets of semantic segmentation We evaluate over three different adversarial targets: "Remapping", "Stripe", and "ICCV 2019". "Remapping" means we generate an adversarial target by shifting the numerical label of each class for the prediction from a benign frame by a constant offset. This way, we can guarantee that each target has no overlap with the ground truth. This dynamically generated target also reflects the movement of pixels between frames in a video. For "Stripe", we divide the target into 19 strips evenly, each of which is filled with a class label, aiming to mitigate possible bias for different classes. Finally, "ICCV 2019" places the text "ICCV 2019" over the image with contiguous spaces representing different classes.

Bounding box mIoU We use a metric called bounding box mIoU as the consistency metric for object detection model. Given the target frame $\boldsymbol{X}_t$ and a pseudo frame $\hat{\boldsymbol{X}}_{s\rightarrow t}$, we iterate each the bounding box predicted by the detection model from the pseudo frame and compute intersection over union (IoU) value against all the bounding boxes detected in the target frame $\boldsymbol{X}_t$, agnostic of their class labels. The average of the largest IoU value for each bounding box in the pseudo frame is used as our consistency metric. The computation of bounding box mIoU is fully described in Algorithm 1 and the computation of IoU between patches is also illustrated below.

Let a patch be represented by a tuple (x_1, y_1, x_2, y_2) where (x_1, y_1) and (x_2, y_2) are the upper left and lower right corners' coordinates of the patch. Given two patches $P = (x_1, y_1, x_2, y_2)$, $P' = (x_1^{'}, y_1^{'}, x_2^{'}, y_2^{'})$, the intersection area of the two patches $\mathcal{A}$ can be computed by $\mathbf{max}(0, \mathbf{min}(x_2, x_2^{'}) - \mathbf{max}(x_1, x_1^{'})) \cdot \mathbf{max}(0, \mathbf{min}(y_2, y_2^{'}) - \mathbf{max}(y_1, y_1^{'}))$. Let w and h denote the width and height of P and w', h' denote the width and height of P'. $\mathbf{getIoU}(P, P')$ is defined as

$$\frac{\mathcal{A}}{w \cdot h + w' \cdot h' - \mathcal{A}} \tag{1}$$

Consistency measurement function for action recognition In our paper, we leverage the attack method proposed by Wei et al. [5]. The target action recognition framework considered in that work [5] is a CNN+RNN model, so we use the same model to evaluate *AdvIT* . Let $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$ be a video where $\boldsymbol{X}_i$ is the ith frame. Each frame is first fed into a Inception V2 [4] model individually and the extracted features are further processed by a LSTM model sequentially. The LSTM model outputs class scores (logits) $\boldsymbol{Y}_i$ for each frame; the logits of all frames are averaged and the class with the highest score is assigned to $\boldsymbol{X}$. The sparse adversarial perturbations are generated using the same method described in the original work [5]. We consider the activations of all frames in our consistency metric to determine if a whole video clip is adversarial. Let $\hat{\boldsymbol{X}} = \{\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}_2, \ldots, \hat{\boldsymbol{X}}_N\}$ be a sequence of pseudo-frames generated by warping with optical flow and $\hat{\boldsymbol{Y}}_i$ be

[*] This work was performed when Chaowei Xiao was at IBM

```
Algorithm 1: Bounding Box mIoU
input:  array of target bounding boxes $\boldsymbol{T}$;
        array of predicted bounding boxes $\boldsymbol{P}$;
output: mIoU between two arrays of bounding boxes $c$;
Initialization : cs ←[], $N \leftarrow \boldsymbol{T}.length, M \leftarrow \boldsymbol{P}.length$;
for $i \leftarrow 1$ to $N$ do
    IoU ← [];
    /* Iterate over the bounding boxes in P. Each bounding box is represented as a patch with its coordinates. */;
    B ← $\boldsymbol{P}[i]$;
    for $j \leftarrow 1$ to $M$ do
        /* Iterate over the bounding boxes in T. */;
        B′ ← $\boldsymbol{T}[j]$;
        /* getIoU is a function that calculate the IoU between two patches. */;
        IoU $\overset{+}{\leftarrow}$ getIoU(B, B′);
    end
    /* Get the IoU of the bounding box in T with the largest overlap with B */;
    cs $\overset{+}{\leftarrow}$ Max(IoU);
end
$c \leftarrow$ Mean(cs);
Return: c
```

the logit of the ith frame after replacing $\boldsymbol{X}$ with $\hat{\boldsymbol{X}}$. We concatenate all the logits $\boldsymbol{Y}_i$ ($\hat{\boldsymbol{Y}}_i$ for pseudo frames) together to form a one-dimensional vector with $N \cdot C$ elements where C is the class number. Let $\boldsymbol{F}$ and $\hat{\boldsymbol{F}}$ be two one-dimensional vectors by taking the softmax of the concatenations of $\boldsymbol{Y}_i$s and $\hat{\boldsymbol{Y}}_i$s respectively. $\boldsymbol{F}$ and $\hat{\boldsymbol{F}}$ represent two $N \cdot C$-way categorical distributions, $\boldsymbol{P}_F$ and $\boldsymbol{P}_{\hat{F}}$. We take the average of the forward and backward KL divergence between the two distributions, $KL\left(\boldsymbol{P}_F || \boldsymbol{P}_{\hat{F}}\right) + KL\left(\boldsymbol{P}_{\hat{F}} || \boldsymbol{P}_F\right)$, as the consistency metric between $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$.

Adaptive attack algorithm Let's use $\boldsymbol{R}$[1] to denote a flow estimator. It takes two frames $\boldsymbol{X}_{t-i}, \boldsymbol{X}_t$ as input, where $\boldsymbol{X}_t$ is the current frame and $\boldsymbol{X}_{t-i}$ is the ith previous frame, and outputs the flow $O_F = (\boldsymbol{\Delta u}, \boldsymbol{\Delta v})$. We formulate the above as $(\boldsymbol{\Delta u}, \boldsymbol{\Delta v}) = \boldsymbol{R}(\boldsymbol{X}_{t-i}, \boldsymbol{X}_t)$. Denote $\mathbf{Y}_t^a$ as the adversarial target for frame t, l as the loss considered by the attack algorithms (e.g. Houdini [2] and DAG [6]) and g as the targeted machine learning model. For the vanilla attack algorithm, they try to generate the adversarial perturbation $\boldsymbol{E}_t$ by optimizing the following objective:

$$l(g(\boldsymbol{X}_t + \boldsymbol{E}_t), \mathbf{Y}_t^a) \tag{2}$$

To perform adaptive attack, the attacker incorporates temporal continuity into the attack. It generates a perturbation for the current frame and the perturbation can fool both current frame and the pseudo frames. The adaptive attack objective is defined as follows:

$$l(g(\boldsymbol{X}_t + \boldsymbol{E}_t), \mathbf{Y}_t^a) + \sum_{i\ =1}^{k} l(\mathbf{warp}(\boldsymbol{R}(\boldsymbol{X}_{t-i}, \boldsymbol{X}_t + \boldsymbol{E}_t) + \boldsymbol{\alpha}, \boldsymbol{X}_{t-i}), \mathbf{Y}_t^a) \tag{3}$$

where k is number of the previous frames considered for adversarial detection. **warp** represents the function to generate the pseudo frames by using formulation (1). $\boldsymbol{\alpha}$ is random noise drawn from $N(0, \sigma^2)$ independently. Due to the randomness, we generate adversarial perturbation using Expectation Over Transformation in Athalye et al. [1]. We follow the settings in Athalye et al. [1] and sample N_z $\boldsymbol{\alpha}$s in each iteration.

$$l(g(\boldsymbol{X}_t + \boldsymbol{E}_t), \mathbf{Y}_t^a) + 1/N_z \cdot \sum_{z=0}^{N_z} \sum_{i=1}^{k} l(\mathbf{warp}(\boldsymbol{R}(\boldsymbol{X}_{t-i}, \boldsymbol{X}_t + \boldsymbol{E}_t) + \boldsymbol{\alpha}_z, \boldsymbol{X}_{t_i}), \mathbf{Y}_t^a) \tag{4}$$

We set $N_z = 30$ in our attack.

[1] In our paper, we use Flownet [3] which is differentiable.

Task	Attack Method	Target	Previous Frames	Detection (k)		
				1	3	5
Semantic Segmentation	Houdini	ICCV	Benign	100%	100%	100%
			Adversarial	100%	100%	100%
		Remapping	Benign	100%	100%	100%
			Adversarial	100%	100%	100%
		Stripe	Benign	100%	100%	100%
			Adversarial	100%	100%	100%
	DAG	ICCV	Benign	100%	100%	100%
			Adversarial	100%	100%	100%
		Remapping	Benign	100%	100%	100%
			Adversarial	100%	100%	100%
		Stripe	Benign	100%	100%	100%
			Adversarial	100%	100%	100%
Human Pose Estimation	Houdini	shuffle	Benign	100%	100%	100%
			Adversarial	100%	100%	100%
		Transpose	Benign	100%	100%	100%
			Adversarial	98%	99%	100%
Object Detection	DAG	all	Benign	100%	100%	100%
			Adversarial	100%	100%	100%
		person	Benign	99%	100%	100 %
			Adversarial	97%	98%	100%

Table A: Detection results (AUC) of *AdvIT* against *independent frame attack* on various video tasks with different attack methods and targets.

B. Experimental Results

We include additional results of *AdvIT* in Table A. It shows that our method can achieve almost 100% detection rate among all settings.

Table B evaluate the accuracy of pseudo-frame prediction in terms of Root Mean Square(RMS) error. We observe that the prediction accuracy does not significantly affect the adversarial detection rate.

Task	Attack Method	Previous Frames	Accuracy[RMS](k)		
			1	3	5
Semantic Segmentation	Houdini	Benign	0.045± 0.006	0.054±0.011	0.059±0.013
		Adversarial	0.045±0.006	0.054±0.011	0.060±0.013
Human Pose Estimation	Houdini	Benign	0.098±0.010	0.114±0.027	0.127±0.039
		Adversarial	0.101±0.010	0.117±0.026	0.130±0.039
Object Detection	DAG	Benign	0.069±0.017	0.090±0.027	0.104±0.032
		Adversarial	0.069±0.017	0.090±0.027	0.104±0.032

Table B: Accuracy of the predicted pseudo-frames among different settings

As long as the prediction accuracy is reasonable, the inconsistency phenomenon is very obvious due to adversarial perturbation. We also show experiments results with different randomness (α) added to the optical flow by varying σ. For semantic segmentation, 0.02 and 0.2 are considered for the value of σ. The corresponding RMS values are 0.088± 0.003 and 0.16 ± 0.006 respectively. 100% detection rates are achieved in both settings. It shows that even we more randomness

Interval (k)	Segmentation (ϵ)			Object detection (ϵ)			Human pose (ϵ)		
	2	16	32	2	16	32	2	16	32
1	100%	100%	100%	98.3%	88.3%	83.5%	98.4%	97.9%	96.3%
3	100%	100%	100%	99.8%	95.5%	91.2%	98.6%	98.4%	96.6%
5	100%	100%	100%	99.9%	97.9%	95.15%	98.7%	98.6%	96.7%

Table C: Detection results (AUC) of differnt attack strength

to (α) during wrapping to make the pipeline more robust against adaptive attack, the detection efficacy of *AdvIT* is not compromised.

In addition, we conducted extra experiments by limiting the perturbation magnitude to 2, 16, 32 pixels (in range of [0,255]) to evaluate the effectiveness of *AdvIT* against the attack with different strength in Table C. 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.

References

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

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

[3] 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.

[4] 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.

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

[6] 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.