DeepSAVA: Sparse Adversarial Videos Attacks and Defenses

APPENDIX

A. Calculating the Gradient of SSIM

The SSIM was first proposed in [1], and is detailed in [2]. Given x and $\hat{x}$ as the local pixels taken from the same location of the same frame in the clean video and adversarial video, respectively, the local similarity between them can be computed on three aspects: structures ($s(x,\hat{x})$), contrasts ($c(x,\hat{x})$), and brightness values ($b(x,\hat{x})$). The local SSIM is formed by these terms [2]:

$$S(x,\hat{x}) = s(x,\hat{x}) \cdot c(x,\hat{x}) \cdot b(x,\hat{x}) = \left(\frac{\sigma_{x\hat{x}} + D_1}{\sigma_x \sigma_{\hat{x}} + D_1}\right) \cdot \left(\frac{2\sigma_x \sigma_{\hat{x}} + D_2}{\sigma_x^2 + \sigma_{\hat{x}}^2 + D_2}\right) \cdot \left(\frac{2\mu_x \mu_{\hat{x}} + D_3}{\mu_x^2 + \mu_{\hat{x}}^2 + D_3}\right), \tag{1}$$

The structural similarity index (SSIM) measure in Equation (2) can be expressed as: [2]

$$\mathrm{SSIM}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{(2\mu_x \mu_{\hat{x}} + C_1)(2\sigma_{x\hat{x}} + C_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2)} \tag{2}$$

The mean of x , the variance of x, and co-variance of x and $\hat{x}$ can be represented as μ_x, σ_x^2 and $\sigma_{x\hat{x}}$. They can be calculated respectively:

$$\begin{aligned} \mu_x &= \frac{1}{N_P}\left(\mathbf{1}^T \cdot \mathbf{x}\right) \\ \sigma_x^2 &= \frac{1}{N_P - 1}(\mathbf{x} - \mu_x)^T (\mathbf{x} - \mu_x) \\ \sigma_{x\hat{x}} &= \frac{1}{N_P - 1}(\mathbf{x} - \mu_x)^T (\hat{\mathbf{x}} - \mu_{\hat{x}}) \end{aligned} \tag{3}$$

Given x and $\hat{x}$ as the local pixels taken from the same location of the same frame in the clean video and adversarial video, respectively, the local similarity between them can be computed on three aspects: structures ($s(x,\hat{x})$), contrasts ($c(x,\hat{x})$), and brightness values ($b(x,\hat{x})$). The local SSIM is formed as [2]:

$$S(x,\hat{x}) = s(x,\hat{x}) \cdot c(x,\hat{x}) \cdot b(x,\hat{x}) = \left(\frac{\sigma_{x\hat{x}} + D_1}{\sigma_x \sigma_{\hat{x}} + D_1}\right) \cdot \left(\frac{2\sigma_x \sigma_{\hat{x}} + D_2}{\sigma_x^2 + \sigma_{\hat{x}}^2 + D_2}\right) \cdot \left(\frac{2\mu_x \mu_{\hat{x}} + D_3}{\mu_x^2 + \mu_{\hat{x}}^2 + D_3}\right), \tag{4}$$

where μ_x and $\mu_{\hat{x}}$ denote means, σ_x and $\sigma_{\hat{x}}$ are standard deviations of x and $\hat{x}$, respectively; $\sigma_{x\hat{x}}$ represents the cross correlation of x and $\hat{x}$ after deleting means; D1, D2, and D3 are weight parameters. For SSIM metric, a value of 1 means that the two images compared are the same. As the SSIM is calculated based on pixel level, it use a sliding window method, which moves pixel by pixel by the window across the whole image. As we use uniform pooling to combine the total SSIM for the whole videos, suppose we have N pixels in the total videos, the SSIM can be represented as:

$$\mathrm{SSIM}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{\sum_{i=1}^{N} \cdot \mathrm{SSIM}(\mathbf{x}_i, \hat{\mathbf{x}}_i)}{N} \tag{5}$$

Models	λ value	FR	ASP(SSIM)
CNN+LSTM	0.8	56.94%	0.0429
	1.0	56.94%	0.0412
	1.5	56.94%	0.0401
I3D	0.8	51.22%	0.0316
	1.0	48.78%	0.0268
	1.5	48.17%	0.0198
Inception-v3	0.8	66.05%	0.0534
	1.0	65.14%	0.0518
	1.5	64.22%	0.0454

TABLE I
THE RESULTS OF DEEPSAVA(WITHOUT BO) ON UCF101 DATASET FOR DIFFERENT λ VALUES.

where x_i and $\hat{x}_i$ are the i-th pixel of each frame in the video. To apply the gradient decent optimisation method described in Section 3, we have to compute the gradient of SSIM with respect to the adversarial video example $\hat{\mathbf{X}}$. As equation (9) shows, to compute $\vec{\nabla}_{\hat{\mathbf{X}}} \mathrm{SSIM}(\mathbf{X}, \hat{\mathbf{X}})$, we only need to calculate the gradient $\vec{\nabla}_{\hat{x}_i} \mathrm{SSIM}(x_i, \hat{x}_i)$. The process is represented as follows. [3] We define four parameters to deduce the derivative of local SSIM:

$$\begin{aligned} M_1 &= 2\mu_x \mu_{\hat{x}} + C_1, \quad M_2 = 2\sigma_{x\hat{x}} + C_2 \\ P_1 &= \mu_x^2 + \mu_{\hat{x}}^2 + C_1, \quad P_2 = \sigma_x^2 + \sigma_{\hat{x}}^2 + C_2 \end{aligned} \tag{6}$$

Therefore, the gradient can be expressed as:

$$\begin{aligned} \nabla_{\hat{x}} \mathrm{SSIM}(x,\hat{x}) = \tfrac{2}{N_P P_1^2 P_2^2} [M_1 P_1 (M_2 x - P_2 \hat{x}) \\ + P_1 P_2 (M_2 - M_1) \mu_x + M_1 M_2 (P_1 - P_2) \mu_{\hat{x}} \end{aligned} \tag{7}$$

B. Effects of λ

To decide the value of λ, we applied the DeepSAVA without BO selection on 200 random selected videos of UCF101 dataset to evaluate the effect of λ. The average success perturbation (ASP) is the average of the SSIM score of perturbation for the adversarial examples that could mislead the model successfully:

$$ASP(SSIM)) = avg(SSIM(V_{adv} - V_{original})),$$

where V_{adv} denotes the generated adversarial video that could successfully mislead the classifier and $V_{original}$ is the original video. The results of applying $\lambda = 0.8, 1.0, 1.5$ on three models are presented in Table I. We can see that the bigger the λ, the lower the FR while the lower the perturbation.While, for the CNN+LSTM model, the fooling rate remains the same across all tested λ values, but the perturbation level is the lowest at λ = 1.5. Thus, we choose $\lambda = 1.5$ for the CNN+LSTM model and $\lambda = 1.0$ for I3D and Inception-v3 model to trade off the performance in terms of the fooling rate and average success perturbation.

C. Generated adversarial videos

The adversarial videos generated by DeepSAVA are demonstrated anonymously by https://www.youtube.com/channel/UCBDswZC2QhBhTOMUFNLchCg

D. Generated adversarial frame for target model: I3D

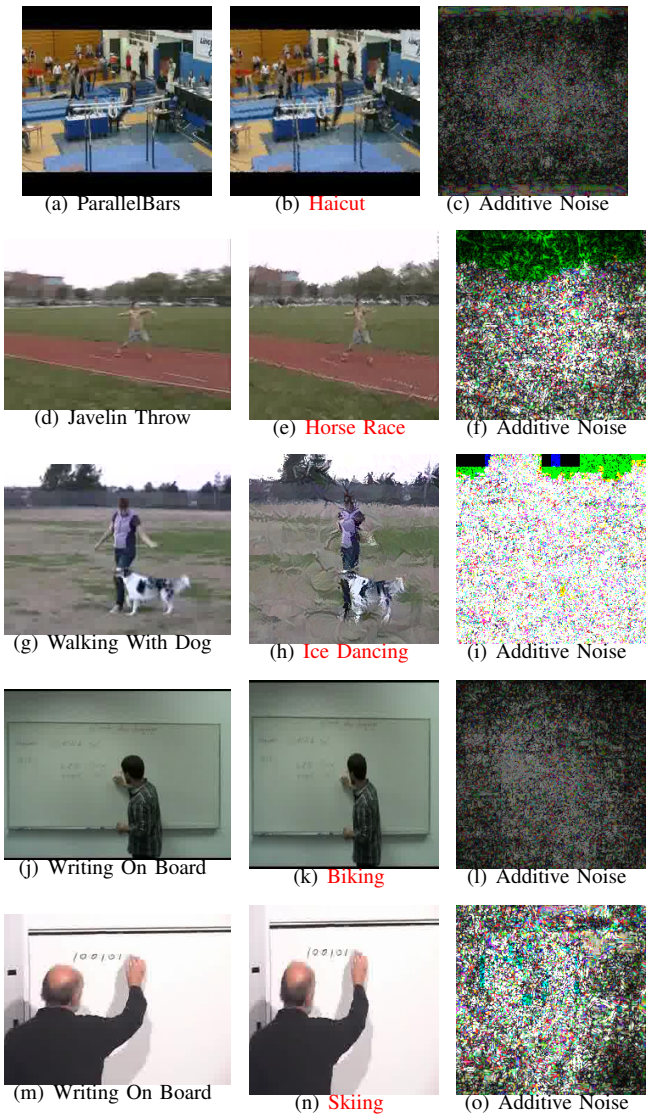

Fig. 1. Original, adversarial examples, additive noise and combined perturbation (normalized into $[0, 1]$ for the visualization) when **only one frame** in the video is perturbed. The red labels are the wrong predictions based on videos for **I3D** model.

E. Generated adversarial frame for target model : CNN+LSTM

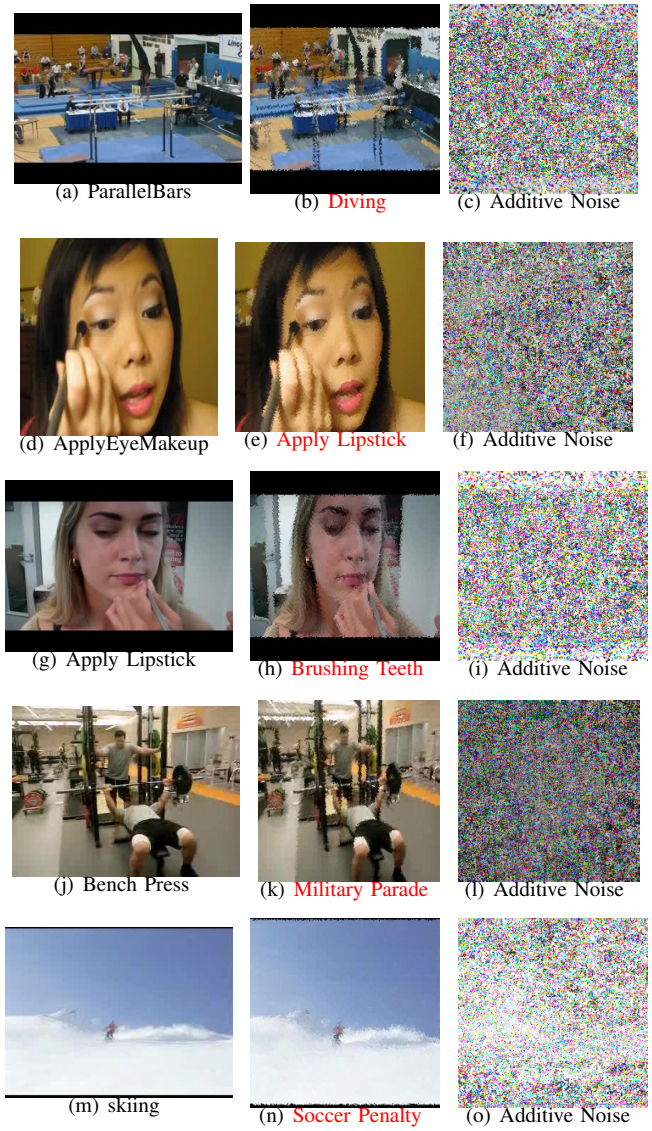

Fig. 2. Original, adversarial examples, additive noise and combined perturbation (normalized into $[0, 1]$ for the visualization) when **only one frame** in the video is perturbed. The red labels are the wrong predictions based on videos for **CNN+LSTM** model.

F. Generated adversarial frame for target model:Inception-v3

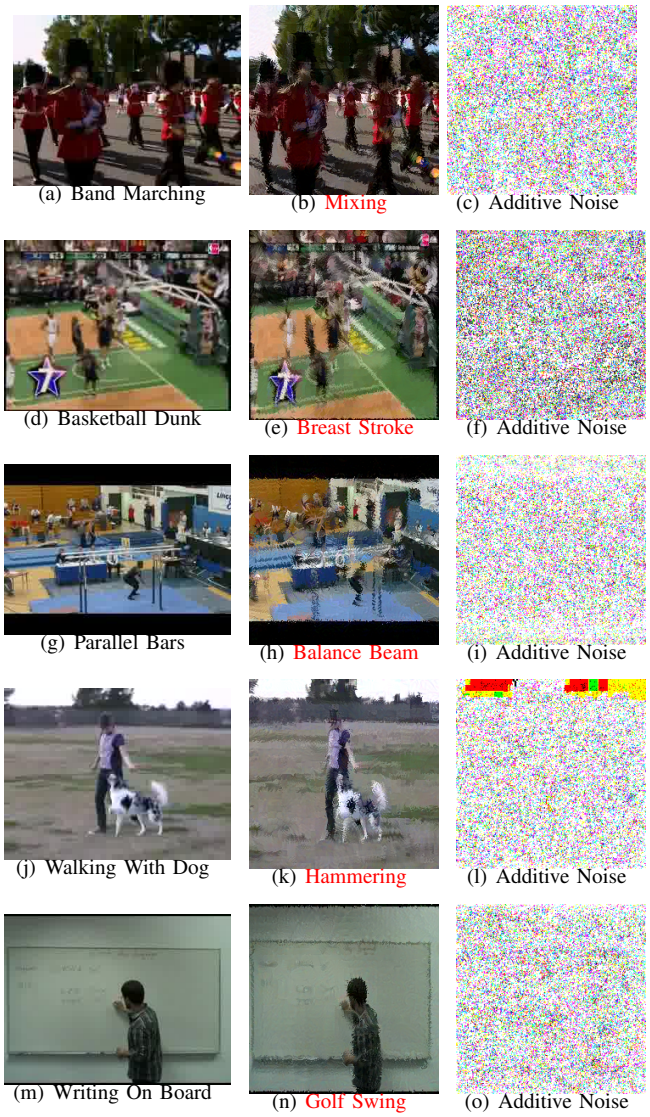

Fig. 3. Original, adversarial examples, additive noise and combined perturbation (normalized into $[0, 1]$ for the visualization) when **only one frame** in the video is perturbed. The red labels are the wrong predictions based on videos for **inception-v3** model.

REFERENCES

[1] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[3] Z. Wang and E. P. Simoncelli, "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics," in *Human Vision and Electronic Imaging IX*, vol. 5292. International Society for Optics and Photonics, 2004, pp. 99–108.