Overcomplete Representations against Adversarial Videos

Shao-Yuan Lo, Jeya Maria Jose Valanarasu and Vishal M.Patel

Johns Hopkins University | Whiting School of Engineering | Baltimore, MD

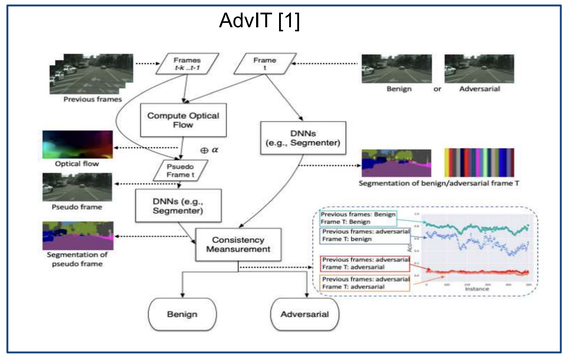
2021 IEEE International Conference on Image Processing
19-22 September 2021 • Anchorage, Alaska, USA
Imaging Without Borders

Introduction

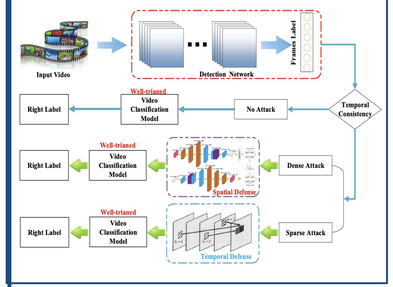
- Adversarial robustness of deep neural networks is an extensively studied problem. However, defenses in the video domain are less explored.
- We propose a novel Over-and-Under complete restoration network for Defending against adversarial videos (OUDefend).
- OUDefend is designed to balance local and global features by learning those two representations.
- OUDefend is attached to a target video recognition model as a feature restoration block and the entire network is trained end-to-end.
- Experimental results show that the defenses focusing on images may be ineffective to videos, while OUDefend enhances robustness against different types of adversarial videos, ranging from additive attacks, multiplicative attacks to physically realizable attacks.

Prior Works

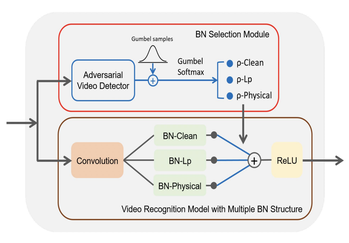
AdvIT [1]



Jia et al. [2]

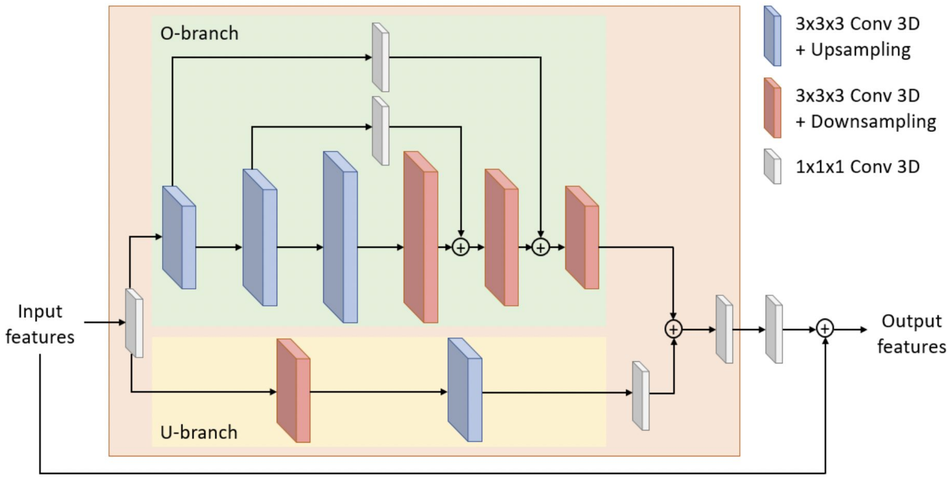


Lo and Patel. [3]



Proposed Method

Architecture of OUDefend



- Most restoration networks adopt an encoder-decoder architecture that first shrinks spatial dimension then expands it back collecting global information but overlooking local details.
- In our method, we propose two branches: an overcomplete branch (Obranch) and an undercomplete branch (U-branch).
- In the encoder of O-branch, each convolutional layer is followed by an upsampling layer, whereas in the decoder, each convolutional layer is followed by a downsampling layer
- U-branch is a standard encoder-decoder structure with downsampling in the encoder and upsampling in the decoder.
- O-branch learns to extract fine details while U-branch learns to extract global information.

References

1. C. Xiao, R. Deng, B. Li, T. Lee, B. Edwards, J. Yi, D. Song, M. Liu, and I. Molloy, “Advit: Adversarial frames identifier based on temporal consistency in videos,” in IEEE International Conference on Computer Vision, 2019.
2. X. Jia, X. Wei, and X. Cao, “Identifying and resisting adversarial videos using temporal consistency,” arXiv preprint arXiv:1909.04837, 2019
3. S.-Y. Lo and V. M. Patel, “Defending against multiple and unforeseen adversarial videos,” arXiv preprint arXiv:2009.05244, 2020.

Results

Experimental Settings

- PGD-ℓ_∞: $\epsilon = 4/255$, $\alpha = 1/255$, and $T = 5$.
- PGD-ℓ_2: $\epsilon = 160$, $\alpha = 1.0$, and $T = 5$.
- MultAV-ℓ_∞: $\epsilon_m = 1.04$, $\alpha_m = 1.01$, and $T = 5$.
- ROA: Rectangle size 30×30, $\epsilon = 255/255$, $\alpha = 70/255$, and $T = 5$.
- AF: Framing width 10, $\epsilon = 255/255$, $\alpha = 70/255$, and $T = 5$.
- SPA: 100 adversarial pixels on each video frame, $\epsilon = 255/255$, $\alpha = 70/255$, and $T = 5$.

Quantitative Results

Method	Params	Clean	PGD-ℓ_∞	PGD-ℓ_2	MultAV	ROA	AF	SPA	Avg$_{adv}$
Clean Model	33.0M	76.90	2.56	3.25	7.19	0.16	0.24	4.39	2.97
Madry’s method [7]	33.0M	76.90	33.94	35.05	47.00	41.29	74.81	55.99	48.01
Xie’s method-A [9]	33.7M	70.82	31.48	33.25	42.69	37.59	58.87	49.14	42.17
Xie’s method-B [9]	34.8M	69.47	30.19	32.65	41.87	38.22	58.74	49.14	41.80
OUDefend (ours)	33.6M	**77.90**	**34.18**	**35.32**	**47.63**	**42.00**	**81.76**	**56.25**	**49.52**

Feature Map Comparison

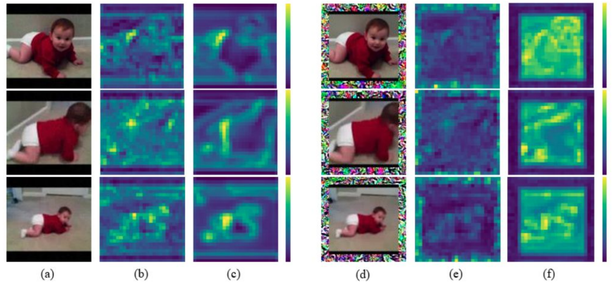


Fig. 2. Feature maps after the conv2 block of Clean Model and OUDefend under PGD-ℓ_∞ and AF. Clean Model is vanilla 3D ResNet-18 trained on clean data. OUDefend is adversarially trained, and here it is inserted after the conv2 block. Top to bottom: Three selected frames from a video. (a) PGD-ℓ_∞ example. (b) Clean Model’s features under PGD-ℓ_∞. (c) OUDefend’s features under PGD-ℓ_∞. (d) AF example. (e) Clean Model’s features under AF. (f) OUDefend’s features under AF.

Acknowledgement

This work was supported by the DARPA GARD Program HR001119S0026-GARD-FP-052

Code

https://github.com/shaoyuanlo/OUDefend