Adversarial Attacks & Defenses in Video

Shao-Yuan Lo
Johns Hopkins University

CVPR 2021 Tutorial on
Adversarial Machine Learning in Computer Vision

June 19, 2021

Outline

- **Image-based** Adversarial Attacks in Video
  - Attacks
  - **Image-based** Defenses
  - **Video-specific** Defenses
- **Video-specific** Adversarial Attacks
- Conclusion

Outline

- **Image-based** Adversarial Attacks in Video
  - Attacks
  - **Image-based** Defenses
  - **Video-specific** Defenses
- **Video-specific** Adversarial Attacks
- Conclusion

Adversarial Attacks in Image

- FGSM [Goodfellow et al. ICLR’15]
- C&W [Carlini et al. SP’17]
- PGD [Madry et al. ICLR’18]

- Adversarial Patch [Brown et al. NeurIPSW’17]
- Rectangular Occlusion Attack (ROA) [Wu et al. ICLR’20]

- A lot more…

Image-based Adversarial Attacks in Video

- Video is a stack of consecutive images.
- A naïve way to generate adversarial videos:
  Use image-based method directly.

$$x^{adv} = x + \epsilon \cdot sign(\nabla_x L(x, y; \theta))$$

$$Image\colon\ x \in R^{C\times H\times W}$$

$$Video\colon\ x \in R^{\boldsymbol{F}\times C\times H\times W}$$

Adversarial Framing (AF)

correct: Boston bull
unattacked: Boston bull
attacked: maypole

correct: ocarina
unattacked: loupe
attacked: maypole

correct: tusker
unattacked: tusker
attacked: maypole

correct: gas pump
unattacked: gas pump
attacked: maypole

correct: Egyptian cat
unattacked: tabby
attacked: maypole

Task: Action recognition
Dataset: UCF-101

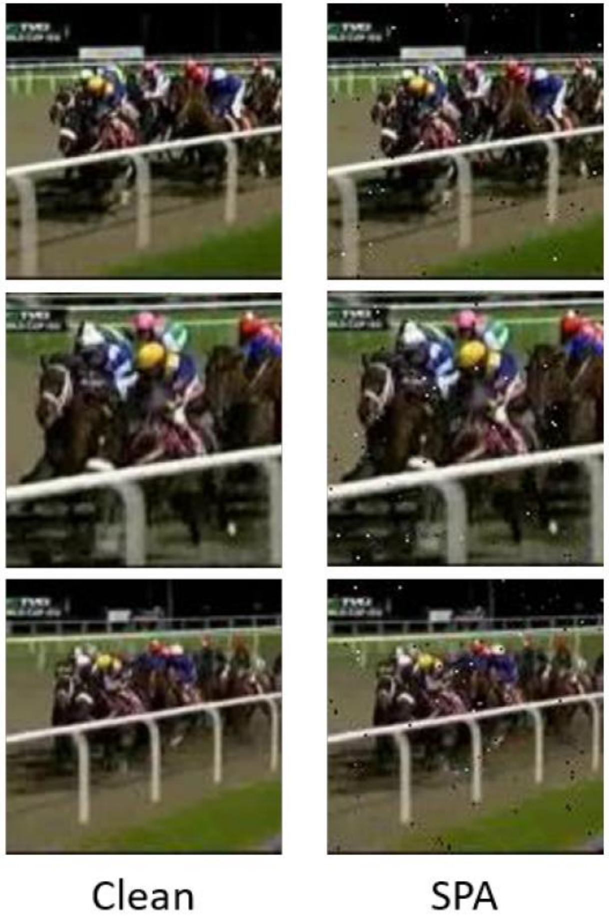
Attack	$W = 1$	$W = 2$	$W = 3$	$W = 4$
None	85.95%			
RF	82.57%	80.53%	81.11%	79.74%
BF	84.94%	84.73%	84.75%	84.59%
AF	65.77%	22.12%	9.45%	2.05%

Michał Zajac, Konrad Zołna, Negar Rostamzadeh, and Pedro O Pinheiro. Adversarial framing for image and video classification. AAAI 2019.

Salt-and-Pepper Attack (SPA)

- Add unbounded perturbations on a number of randomly selected pixels.
- The perturbation looks like salt-and-pepper noise.
- A kind of L0 attack.

- Decrease action recognition accuracy from **89.0%** to **8.4%** on UCF-101.



Shao-Yuan Lo and Vishal M. Patel. Defending Against Multiple and Unforeseen Adversarial Videos. 2020.

Multiplicative Adversarial Videos (MultAV)

- Additive:

$$\mathbf{x}^{t+1} = Clip^{\ell_\infty}_{\mathbf{x},\epsilon}\{\mathbf{x}^t + \alpha \cdot sign(\nabla_{\mathbf{x}^t}\mathcal{L}(\mathbf{x}^t, \mathbf{y};\ \boldsymbol{\theta}))\}$$

$$\mathbf{x}^{t+1} = Clip^{\ell_2}_{\mathbf{x},\epsilon}\{\mathbf{x}^t + \alpha \cdot \frac{\nabla_{\mathbf{x}^t}\mathcal{L}(\mathbf{x}^t, \mathbf{y};\ \boldsymbol{\theta})}{\| \nabla_{\mathbf{x}^t}\ \mathcal{L}(\mathbf{x}^t, \mathbf{y};\ \boldsymbol{\theta})\|_2}\}$$

- Multiplicative:

$$\mathbf{x}^{t+1} = Clip^{RB-\ell_\infty}_{\mathbf{x},\epsilon_m}\{\mathbf{x}^t \odot \alpha_m^{sign(\nabla_{\mathbf{x}^t}\mathcal{L}(\mathbf{x}^t, \mathbf{y};\boldsymbol{\theta}))}\}$$

$$\mathbf{x}^{t+1} = Clip^{RB-\ell_2}_{\mathbf{x},\epsilon_m}\{\mathbf{x}^t \odot \alpha_m^{\frac{\nabla_{\mathbf{x}^t}\mathcal{L}(\mathbf{x}^t, \mathbf{y};\boldsymbol{\theta})}{\|\nabla_{\mathbf{x}^t}\mathcal{L}(\mathbf{x}^t, \mathbf{y};\boldsymbol{\theta})\|_2}}\}$$

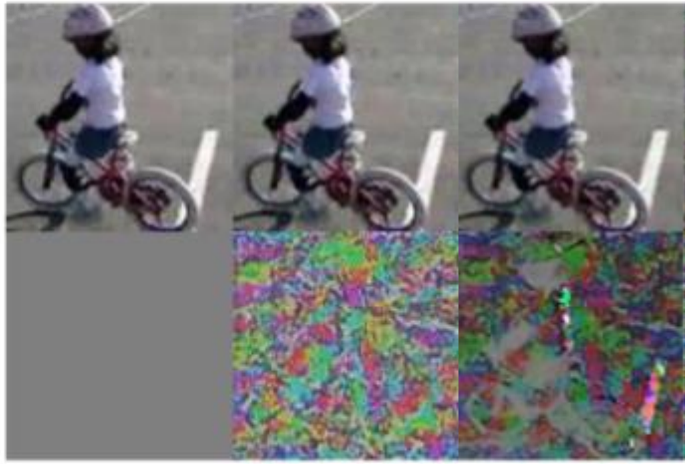
Shao-Yuan Lo and Vishal M. Patel. MultAV: Multiplicative Adversarial Videos. 2020.

Multiplicative Adversarial Videos (MultAV)

Task: Action recognition
Dataset: UCF-101

Network	Clean
3D ResNet-18	76.90

MultAV-ℓ_∞	MultAV-ℓ_2	MultAV-ROA	MultAV-AF	MultAV-SPA
7.19	2.67	2.30	0.26	4.02



Clean PGD-ℓ_∞ MultAV-ℓ_∞

Shao-Yuan Lo and Vishal M. Patel. MultAV: Multiplicative Adversarial Videos. 2020.

Outline

- **Image-based** Adversarial Attacks in Video
  - Attacks
  - **Image-based** Defenses
  - **Video-specific** Defenses
- **Video-specific** Adversarial Attacks
- Conclusion

Adversarial Training in Video

- Adversarial Training (AT) is considered one of the most effective defenses, especially in the white-box setting.
- Madry et al. [ICLR'18] formulated AT in a min-max optimization framework:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{D}} \left[\max_{\delta \in \mathbb{S}} L(x+\delta, y; \theta) \right]$$

$$Image\colon\ x \in R^{C\times H\times W}$$

$$Video\colon\ x \in R^{\boldsymbol{F}\times C\times H\times W}$$

AT Benchmark in Video

- Dataset: UCF-101 (action recognition)
- Model: 3D ResNet-18 (**76.90%** clean accuracy)
- Attacks:
  - PGD Linf: ε=4/255, T=5
  - PGD L2: ε=160, T=5
  - MultAV: ε=1.04, T=5
  - ROA: patch size=30x30, T=5
  - SPA: # pixels=100, T=5

Method	**PGD Linf**	**PGD L2**	**MultAV**	**ROA**	**SPA**
No Defense	2.56	3.25	7.19	0.16	4.39
AT	33.94	35.05	47.00	41.29	55.99

Shao-Yuan Lo, Jeya Maria Jose Valanarasu, and Vishal M. Patel. Overcomplete Representations Against Adversarial Videos. ICIP 2021.

AT Benchmark in Video

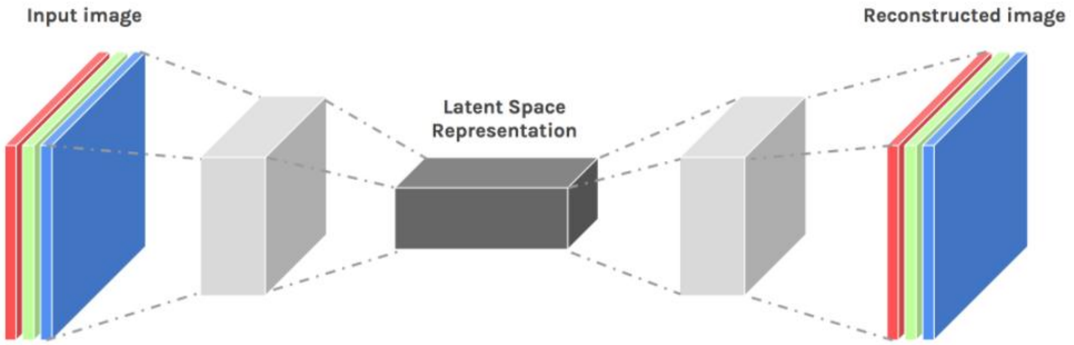
- Dataset: UCF-101 (action recognition)
- Model: 3D ResNeXt-101 (**89.0%** clean accuracy)
- Attacks:
  - PGD Linf: ε=4/255, T=5
  - ROA: patch size=30x30
  - AF: width=10
  - SPA: #pixels=100, T=5

Method	PGD Linf	ROA	AF	SPA
No Defense	3.3	0.5	1.6	8.3
AT	49.0	69.0	80.5	60.4

Shao-Yuan Lo and Vishal M. Patel. Defending Against Multiple and Unforeseen Adversarial Videos. 2020.

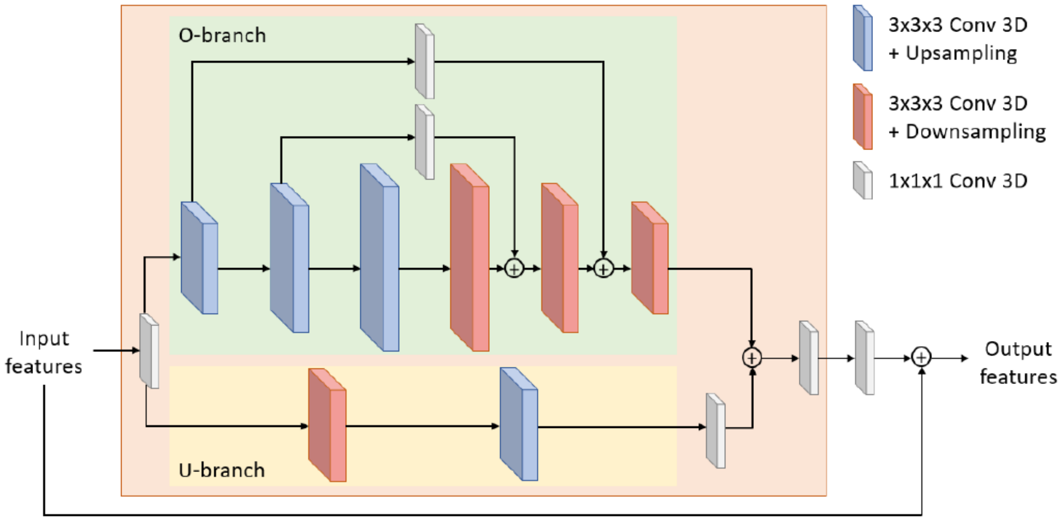
Overcomplete Representations Against Adversarial Videos (OUDefend)

- A typical autoencoder downsample features and learn **undercomplete** representations.



https://ai.plainenglish.io/convolutional-autoencoders-cae-with-tensorflow-97e8d8859cbe.

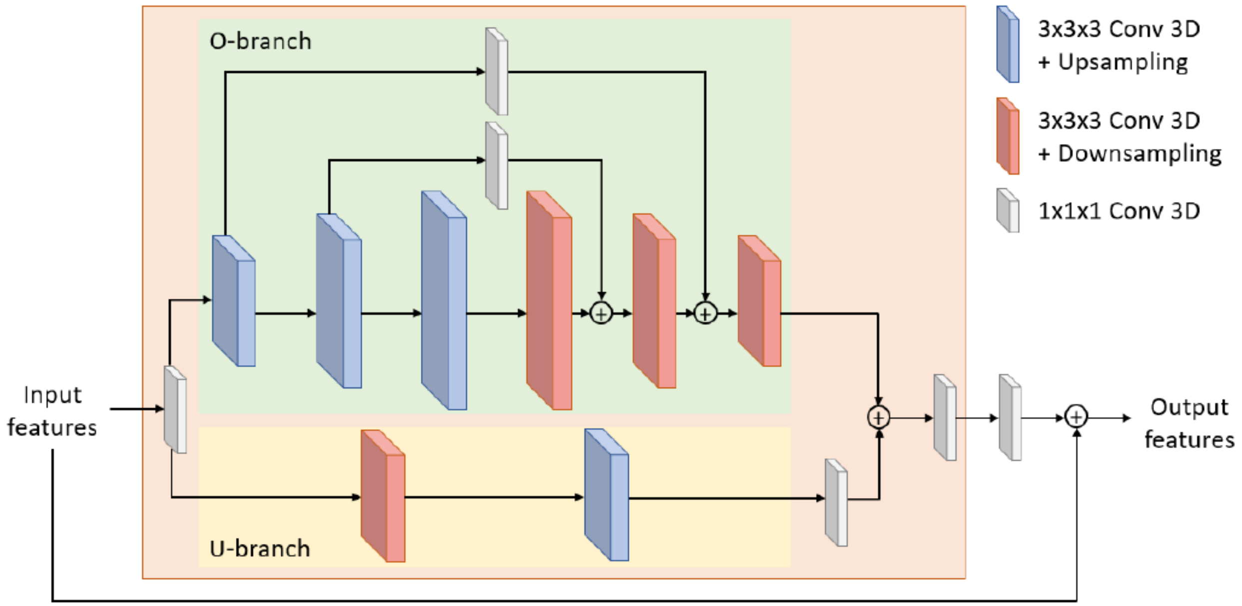
- OUDefend learns both **undercomplete** representations and **overcomplete** representations (upsample features)



Shao-Yuan Lo, Jeya Maria Jose Valanarasu, and Vishal M. Patel. Overcomplete Representations Against Adversarial Videos. ICIP 2021.

Overcomplete Representations Against Adversarial Videos (OUDefend)

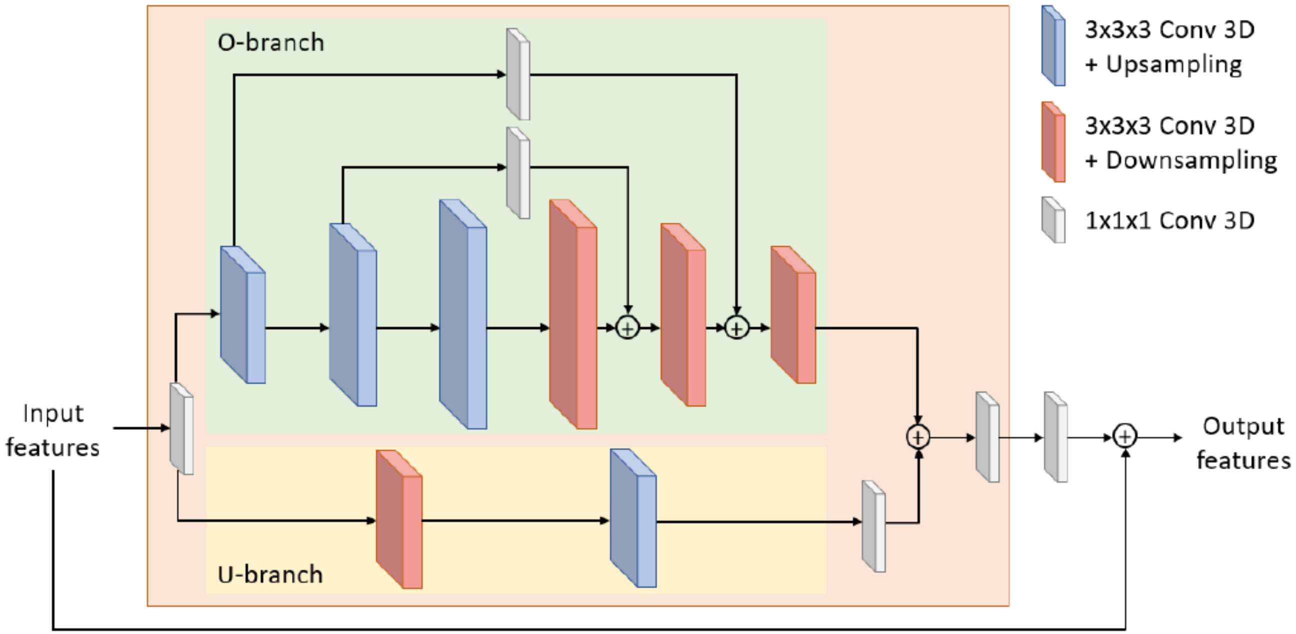
- **Undercomplete** representations have large receptive fields to collect global information, but it overlooks local details.
- **Overcomplete** representations have opposite properties.
- OUDefend balances **local** and **global** features by learning those two representations.



Shao-Yuan Lo, Jeya Maria Jose Valanarasu, and Vishal M. Patel. Overcomplete Representations Against Adversarial Videos. ICIP 2021.

Overcomplete Representations Against Adversarial Videos (OUDefend)

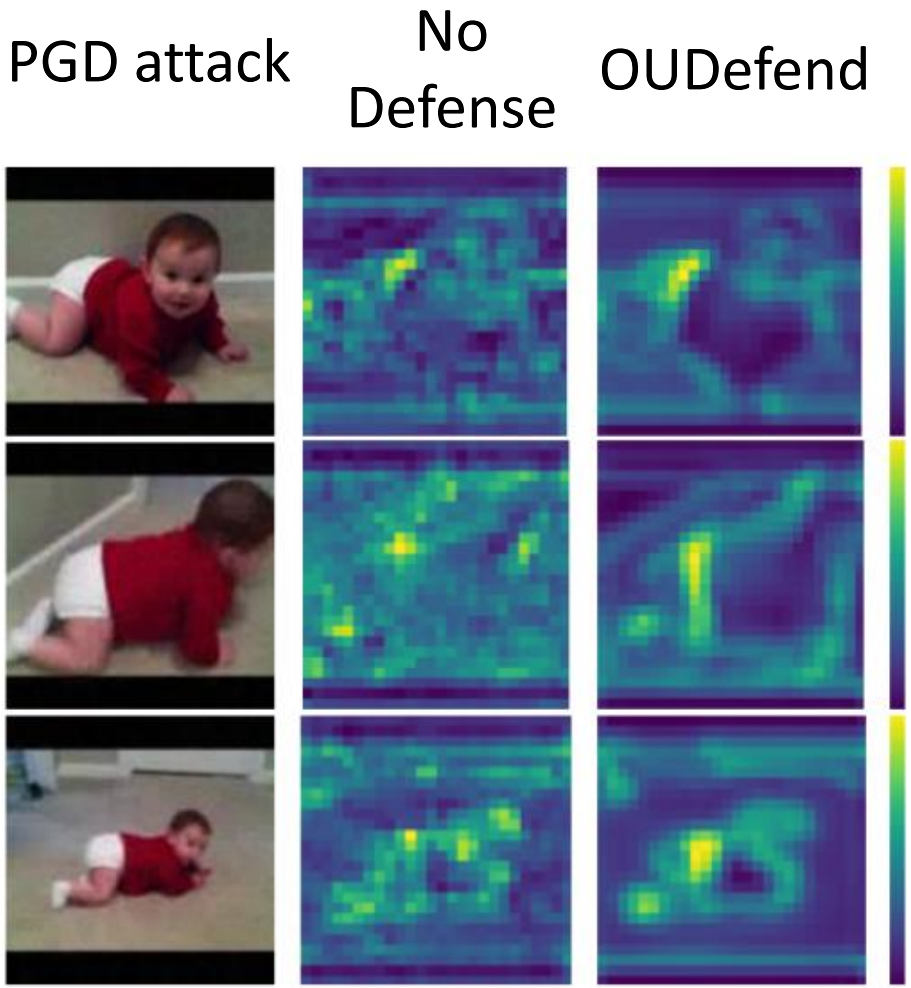
- Append OUDefend blocks to the target network (after each res block).



layer name	output size	18-layer
conv1	112×112	
conv2_x	56×56	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$

Shao-Yuan Lo, Jeya Maria Jose Valanarasu, and Vishal M. Patel. Overcomplete Representations Against Adversarial Videos. ICIP 2021.

Overcomplete Representations Against Adversarial Videos (OUDefend)



Method	PGD Linf	PGD L2	MultAV	ROA	SPA
No Defense	2.56	3.25	7.19	0.16	4.39
AT	33.94	35.05	47.00	41.29	55.99
OUDefend	**34.18**	**35.32**	**47.63**	**42.00**	**56.29**

Shao-Yuan Lo, Jeya Maria Jose Valanarasu, and Vishal M. Patel. Overcomplete Representations Against Adversarial Videos. ICIP 2021.

Multi-Perturbation Robustness in Video

Clean PGD ROA AF SPA

How to defend against multiple types of attacks simultaneously?

Multi-Perturbation Robustness in Video

- Standard AT has suboptimal multi-perturbation robustness.
- Training: δ_{PGD}
- Test: Clean, δ_{PGD}, δ_{ROA}, δ_{AF}, δ_{SPA}

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{D}} \underbrace{\left[\max_{\delta\in\mathbb{S}} L(x+\delta, y; \theta)\right]}_{\text{Generate one type of adversarial examples}}$$

Generate **one type** of
adversarial examples

Multi-Perturbation Robustness in Video

- Average AT is better, but not enough.
- Training: Clean, δ_{PGD}, δ_{ROA}, δ_{AF}, δ_{SPA}
- Test: Clean, δ_{PGD}, δ_{ROA}, δ_{AF}, δ_{SPA}

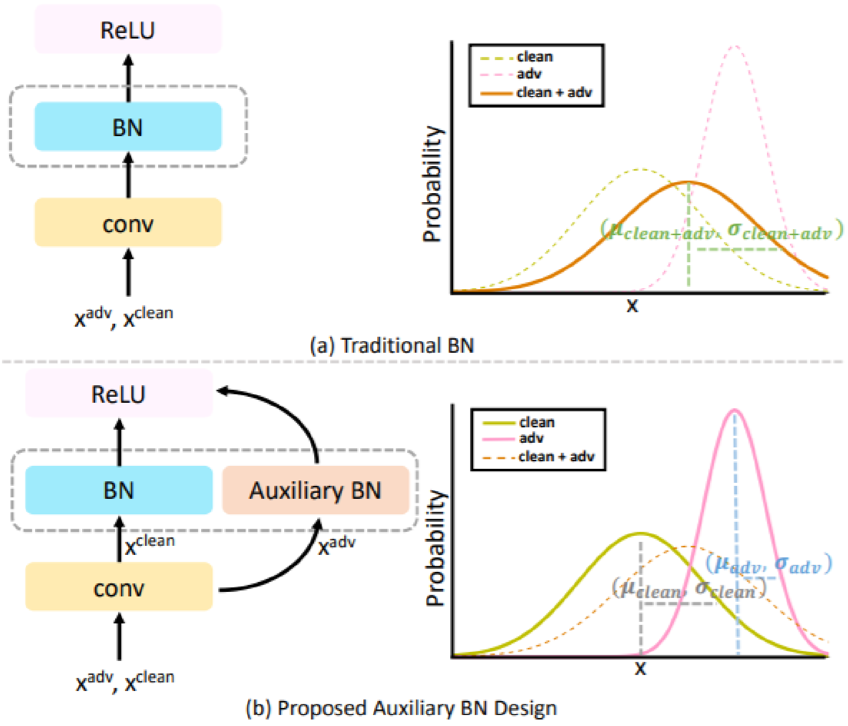
$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{D}} \underbrace{\left[\sum_{i=1}^{N} \max_{\delta_i \in \mathbb{S}_i} L(x + \delta_i, y; \theta) \right]}$$

Generate **multiple types** of adversarial examples

Florian Tramèr and Dan Boneh. Adversarial Training and Robustness for Multiple Perturbations. NeurIPS 2019.

Multi-Perturbation Robustness in Video

- Why is average AT **not** an ideal strategy?
- Example: **Clean** vs. **PGD**.
- Clean and PGD have distinct data distributions.
- The statistics estimation at BN may be confused when facing a mixture distribution.
- An auxiliary BN guarantees that data from different distributions are normalized separately.



Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan Yuille, Quoc Le. Adversarial Examples Improve Image Recognition. CVPR 2020.

Multi-Perturbation Robustness in Video

- What about **multiple** attack types?
- Example: Clean, PGD, ROA, AF, SPA
- Assumption: Different attack types have **distinct** data distributions.

Shao-Yuan Lo and Vishal M. Patel. Defending Against Multiple and Unforeseen Adversarial Videos. 2020.

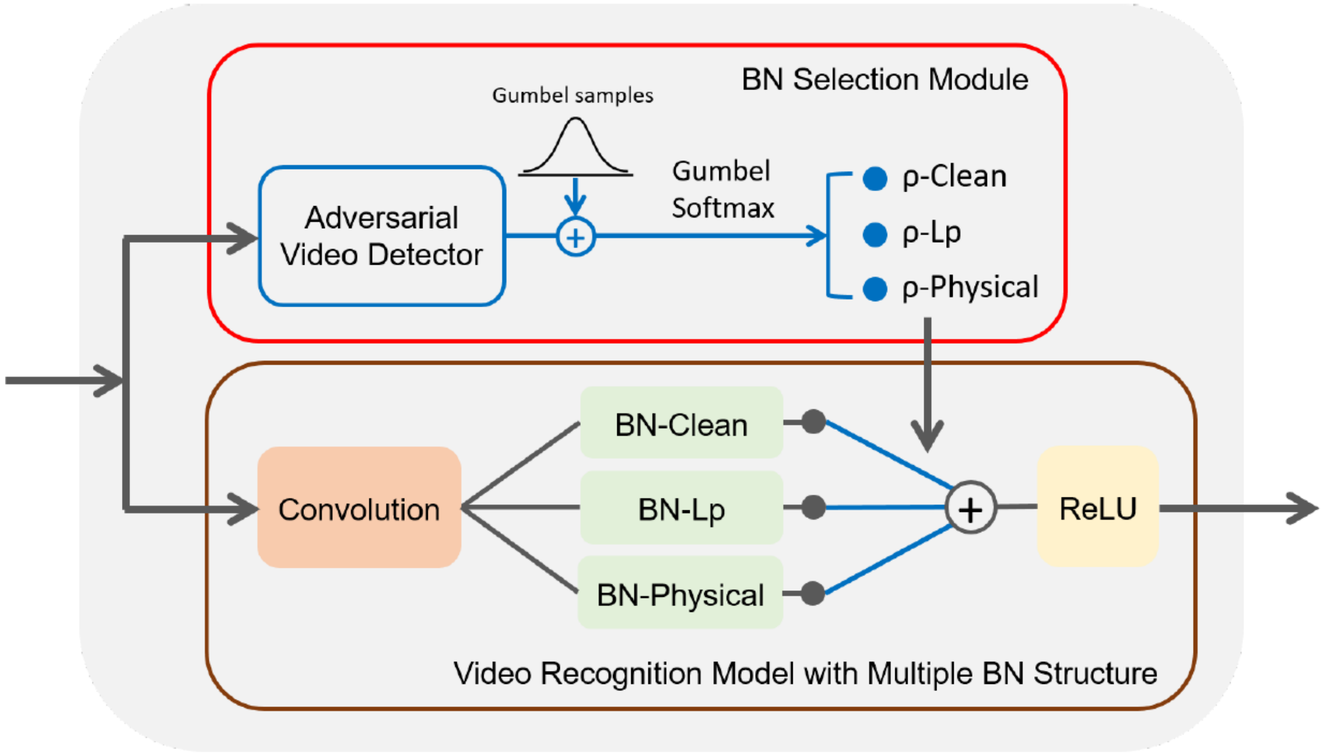
Multi-Perturbation Robustness in Video

- What about **unforeseen** attack types?
- Example:
  - **Known**: Clean, PGD, ROA
  - **Unforeseen**: AF, SPA
- **Digital attacks**: PGD, SPA
- **Physically realizable attacks**: ROA, AF
- Assumption: Similar attack types have **similar** data distributions.

Shao-Yuan Lo and Vishal M. Patel. Defending Against Multiple and Unforeseen Adversarial Videos. 2020.

Multi-Perturbation Robustness in Video

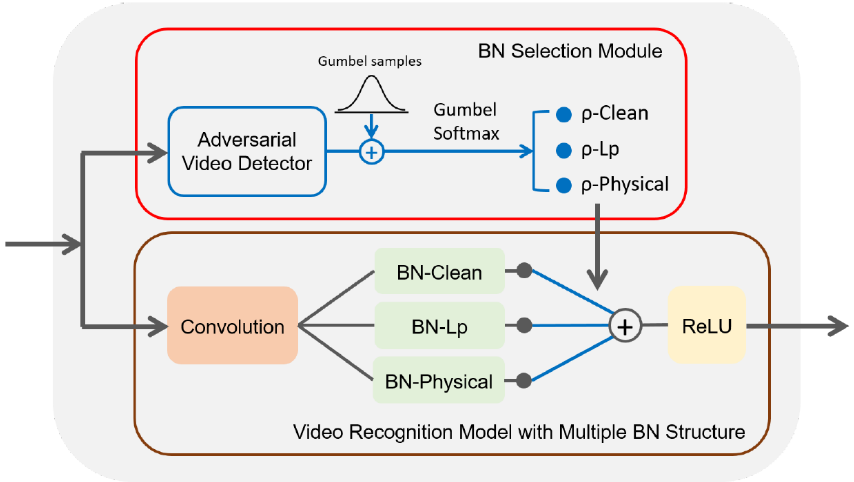
- Example:
  - **Known**: Clean, PGD, ROA
  - **Unforeseen**: AF, SPA
- **Digital attacks**: PGD, SPA
- **Physically realizable attacks**: ROA, AF



Shao-Yuan Lo and Vishal M. Patel. Defending Against Multiple and Unforeseen Adversarial Videos. 2020.

Multi-Perturbation Robustness in Video

- Training: Clean, δ_{PGD}, δ_{ROA}
- Test: Clean, δ_{PGD}, δ_{ROA}, δ_{AF}, δ_{SPA}



$$\theta^* = \arg\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\mathbb{D}} \Big[L(x,y;\theta) + \lambda \cdot L(x, y^{det}; \theta^{det}) + \sum_{i=1}^{N} \big(\max_{\delta_i \in \mathbb{S}_i} L(x+\delta_i, y; \theta) + \lambda \cdot L(x+\delta_i, y^{det}; \theta^{det}) \big) \Big]$$

Shao-Yuan Lo and Vishal M. Patel. Defending Against Multiple and Unforeseen Adversarial Videos. 2020.

Multi-Perturbation Robustness in Video

Dataset: UCF-101

Model	Clean	PGD	ROA	AF	SPA	Mean	Union
No Defense	**89.0**	3.3	0.5	1.6	8.4	20.6	0.0
TRADE [19] (ICML'19)	82.3	29.0	5.7	3.3	42.2	32.5	1.9
AVG [26] (NeurIPS'19)	68.9	38.1	51.4	18.5	49.6	45.3	17.3
MAX [26] (NeurIPS'19)	72.8	32.5	31.0	5.8	49.4	38.3	5.5
MSD [27] (ICML'20)	70.2	43.2	1.7	1.6	**56.0**	34.6	0.7
MultiBN (ours)	74.2	**44.6**	**58.6**	**44.3**	53.7	**55.1**	**34.8**

Dataset: HMDB-51

Model	Clean	PGD	ROA	AF	SPA	Mean	Union
No Defense	**65.1**	0.0	0.0	0.0	0.3	13.1	0.0
TRADE [19] (ICML'19)	54.8	6.8	0.3	0.0	20.5	16.5	0.0
AVG [26] (NeurIPS'19)	39.0	14.3	17.1	2.8	26.2	19.9	1.4
MAX [26] (NeurIPS'19)	48.6	13.9	16.0	0.1	30.3	21.8	0.0
MSD [27] (ICML'20)	41.4	18.2	0.1	0.0	**31.2**	18.2	0.0
MultiBN (ours)	51.1	**22.0**	**23.7**	**7.8**	29.9	**26.9**	**5.0**

Shao-Yuan Lo and Vishal M. Patel. Defending Against Multiple and Unforeseen Adversarial Videos. 2020.

Outline

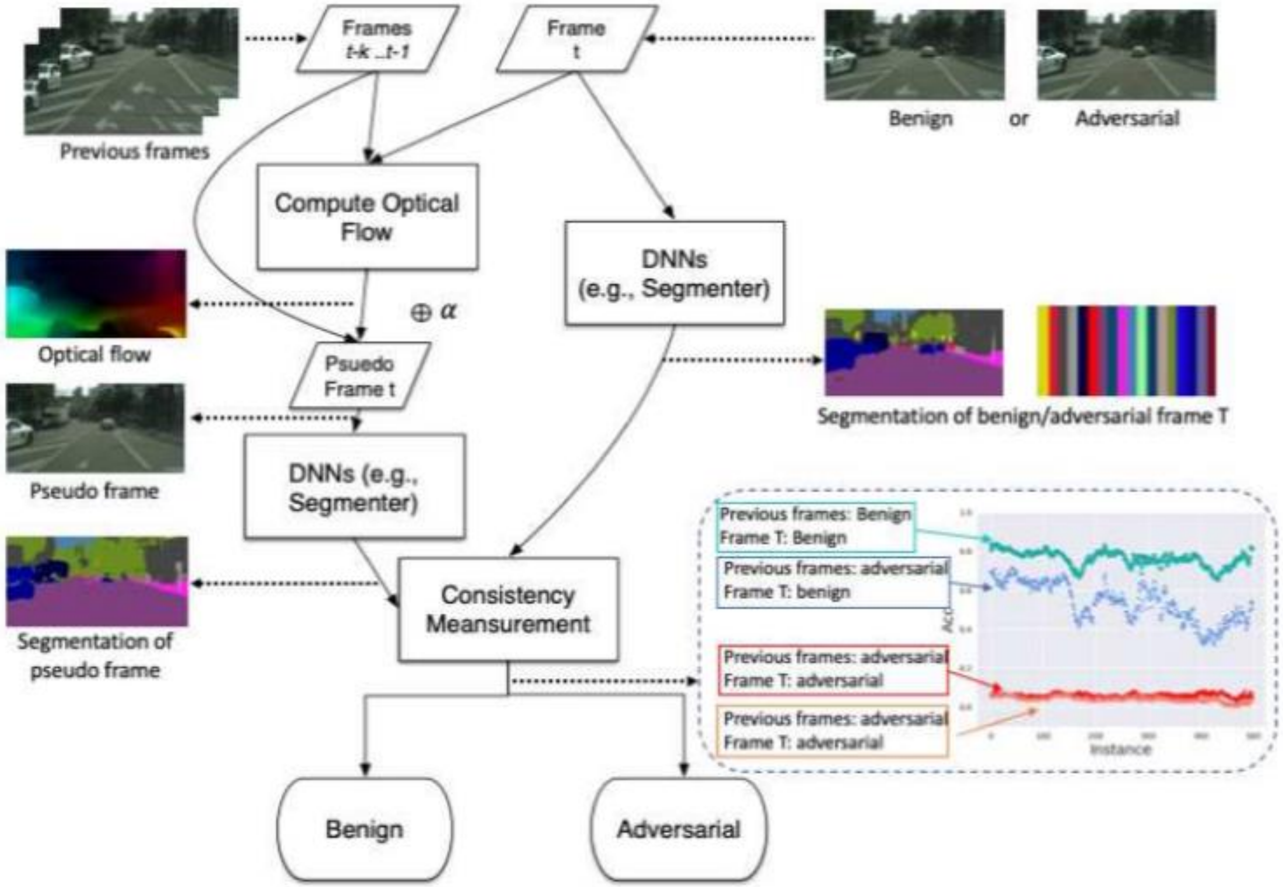
- **Image-based** Adversarial Attacks in Video
  - Attacks
  - **Image-based** Defenses
  - **Video-specific** Defenses
- **Video-specific** Adversarial Attacks
- Conclusion

Video-specific Defenses

- Use video’s unique properties (mostly temporal information) to defend against adversarial videos (image-based attacks).
- Some studies work on adversarial detection.
- Few studies for defense.

AdvIT: Adversarial Frames Identifier Based on Temporal Consistency In Videos

- Compare the output of the target frame and its corresponding pseudo frame.
- The pseudo frame is much less affected by adversary.
- No training.

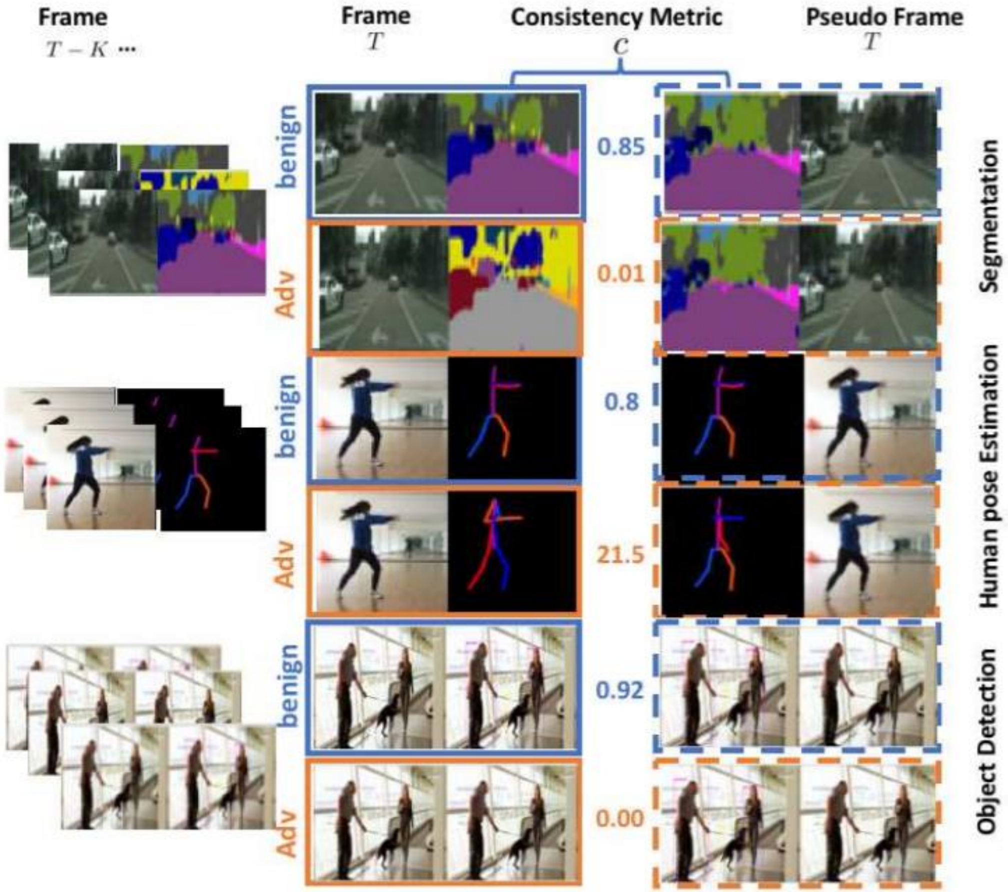


Chaowei Xiao, Ruizhi Deng, Bo Li, Taesung Lee, Jinfeng Yi, Ian Molloy, Mingyan Liu, and Dawn Song. AdvIT: Adversarial Frames Identifier Based on Temporal Consistency In Videos. ICCV 2019.

AdvIT: Adversarial Frames Identifier Based on Temporal Consistency In Videos

- Temporal consistency test

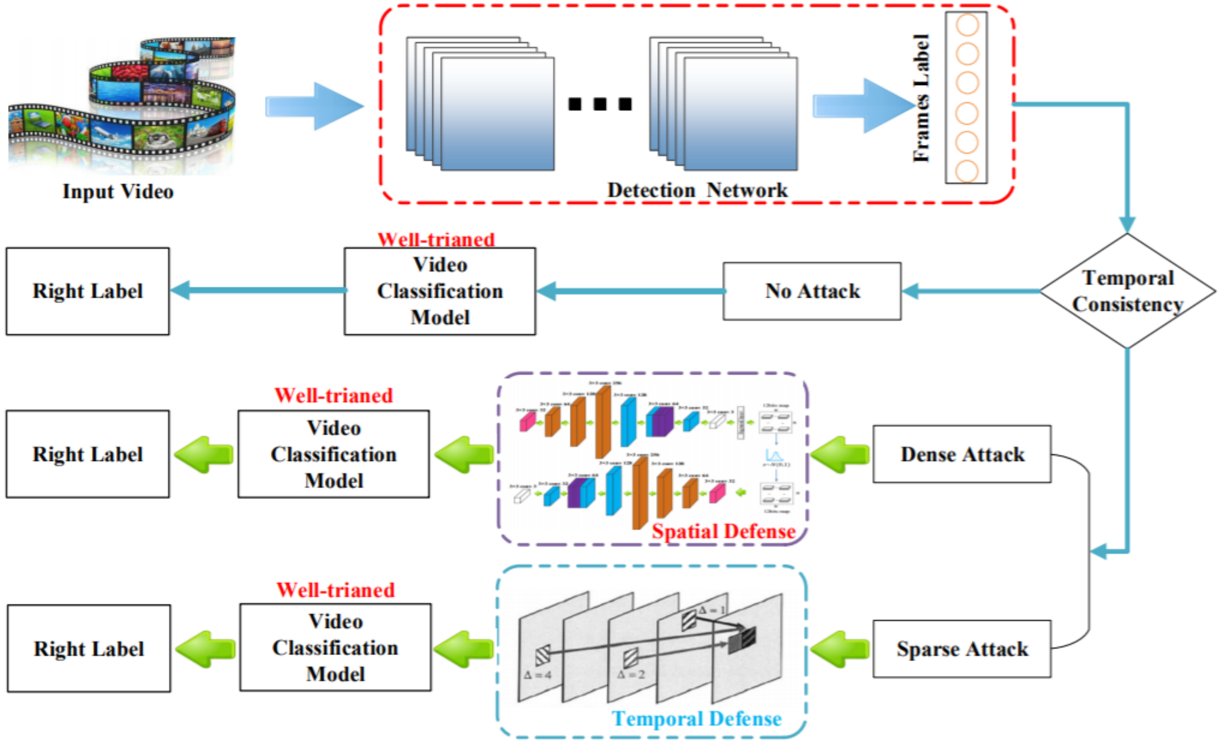
- Semantic segmentation: Pixel-wise accuracy
- Object detection: mIoU of bounding boxes
- Human pose estimation: MSE



Chaowei Xiao, Ruizhi Deng, Bo Li, Taesung Lee, Jinfeng Yi, Ian Molloy, Mingyan Liu, and Dawn Song. AdvIT: Adversarial Frames Identifier Based on Temporal Consistency In Videos. ICCV 2019.

Identifying and Resisting Adversarial Videos Using Temporal Consistency

- Use temporal consistency to detect adversarial frames.
- Spatial Defense: Image-based defense
- Temporal Defense: Replace adversarial frames with pseudo frames



Xiaojun Jia, Xingxing Wei, and Xiaochun Cao. Identifying and Resisting Adversarial Videos Using Temporal Consistency. 2019.

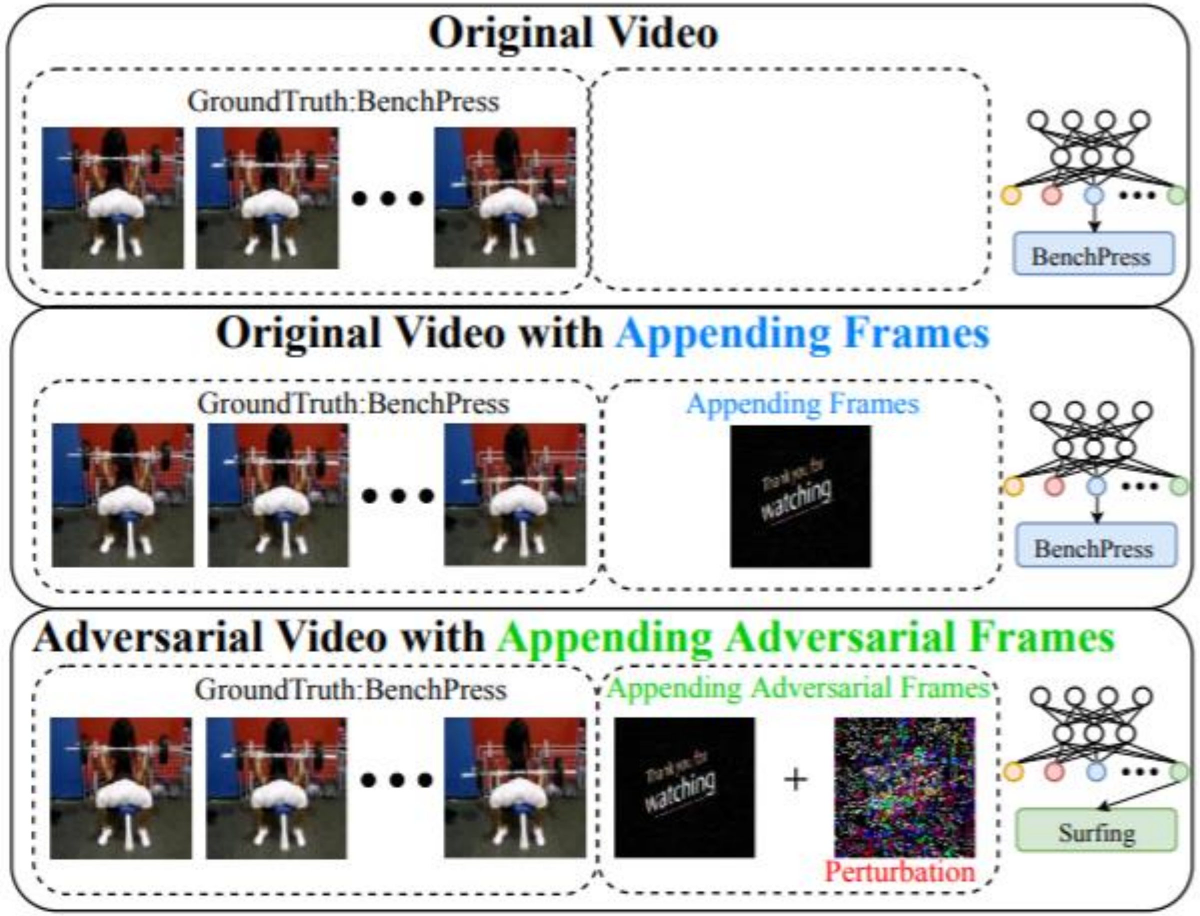
Outline

- **Image-based** Adversarial Attacks in Video
  - Attacks
  - **Image-based** Defenses
  - **Video-specific** Defenses
- **Video-specific** Adversarial Attacks
- Conclusion

Video-specific Defenses

- Use video’s unique properties (mostly temporal information) to generate adversarial videos.
- Video has higher dimensionality, so the search space of adversary is larger -> more possible types of adversarial examples

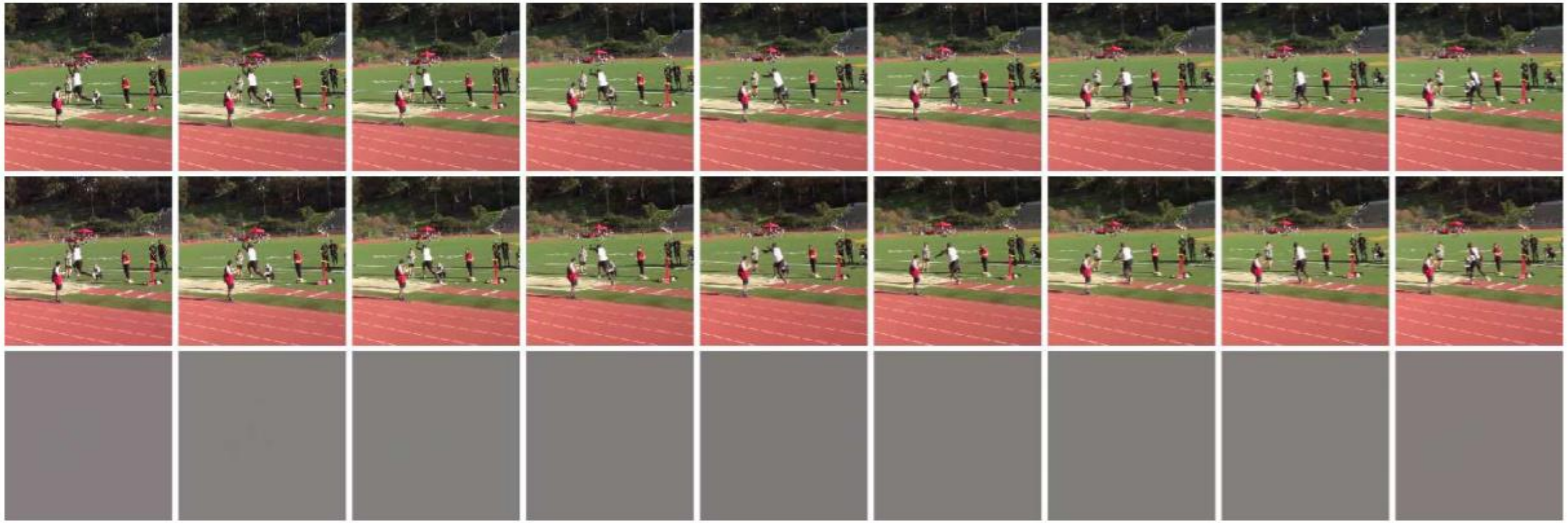
Appending Adversarial Frames



Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Qi Tian. Appending Adversarial Frames for Universal Video Attack. WACV 2021.

Adversarial Flickering Attacks

- Spatial patternless temporal perturbation, i.e., the perturbation is a constant offset applied to the entire frame.
- Undetectable by image adversarial attack detector.



Roi Pony, Itay Naeh, and Shie Mannor. Over-the-Air Adversarial Flickering Attacks against Video Recognition Networks. 2020.

Adversarial Flickering Attacks

- Objective function (universal targeted attack)

$$\operatorname*{argmin}_{\delta} \lambda \sum_{j} \beta_j D_j(\delta) + \frac{1}{N} \sum_{n=1}^{N} \ell(F_\theta(X_n + \delta), t_n)$$

- F_ϑ is classifier
- *N* is total number of training videos
- *t* is targeted class
- D_j is regularization term
- β_j weights the relative importance of each regularization term
- λ weights the relative importance of the regularization terms

Adversarial Flickering Attacks

- Thickness regularization: Force the perturbation to be small.

$$D_1(\delta) = \frac{1}{3T} \|\delta\|_2^2 .$$

- Roughness regularization: Force the perturbation to be smooth.

$$D_2(\delta) = D_2^1(\delta) + D_2^2(\delta)$$

$$D_2^1(\delta) = \frac{1}{3} \sum_{c \in \{r,g,b\}} \frac{1}{T-1} \sum_{i=2}^{T} \left\| \delta_i^c - \delta_{i-1}^c \right\|_2^2$$

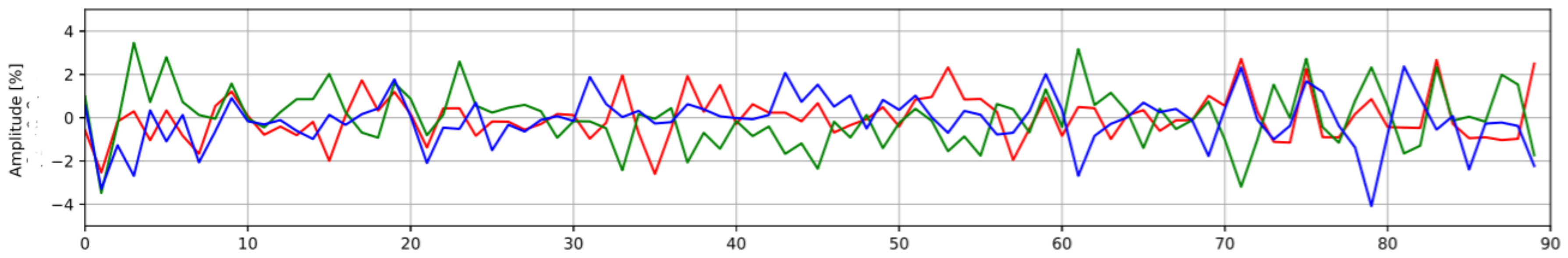
Control the difference between two consecutive frame perturbations

$$D_2^2(\delta) = \frac{1}{3} \sum_{c \in \{r,g,b\}} \frac{1}{T-2} \sum_{i=2}^{T-1} \left\| \delta_{i+1}^c - 2\delta_i^c + \delta_{i-1}^c \right\|_2^2$$

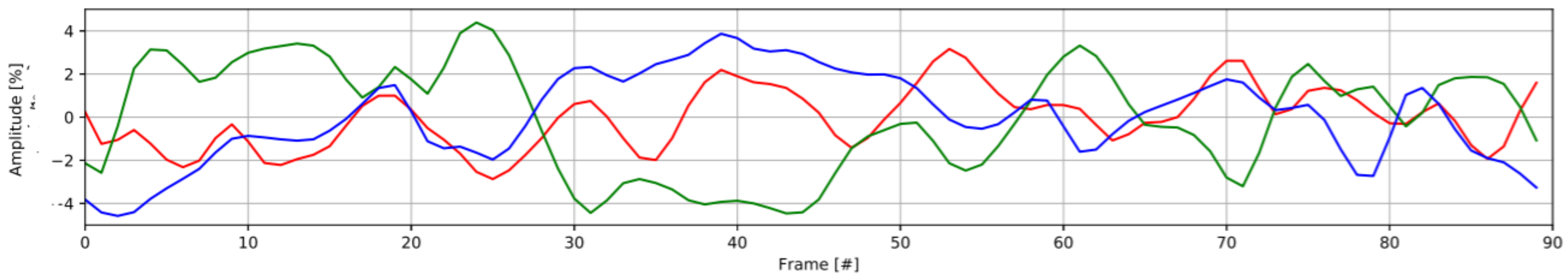
Control the trend of perturbation

Adversarial Flickering Attacks

- Using D1 only



- Using D2 only



Conclusion

- Image-based adversarial attack and defense methods can generalize to video.

- With video-specific properties, there exist more possible types of adversarial videos.

- Video-specific defense is still an open problem.

Thanks for your attention