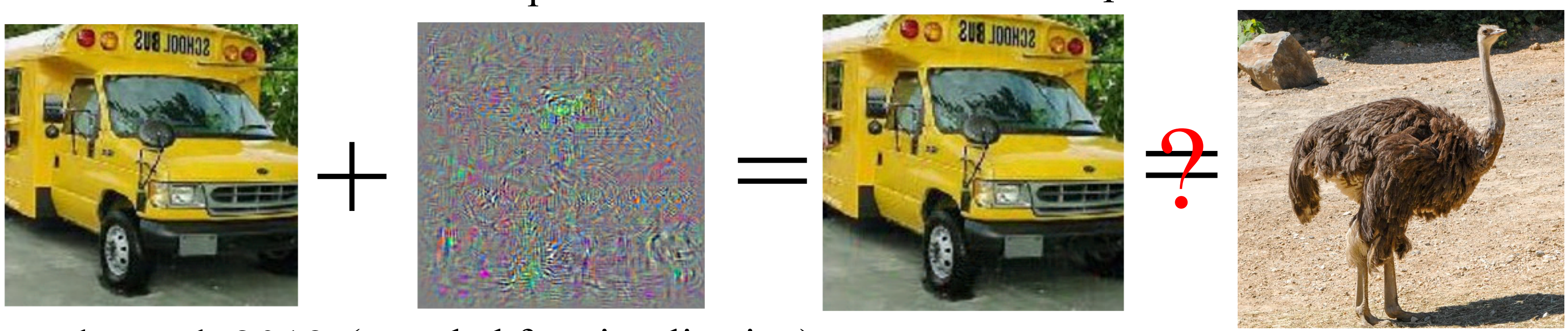
Stealthy Adversarial Perturbations against Real-time Video Classification Systems

Shasha Li*, Ajaya Neupane*, Sujoy Paul*, Chengyu Song*,
Srikanth V. Krishnamurthy*, Amit K. Roy Chowdhury* and Ananthram Swami †
* University of California Riverside
† United States Army Research Laboratory

Adversarial Perturbations

Adversarial Perturbations Against Real-Time Video Classification Systems

"adversarial perturbation" "adversarial example"



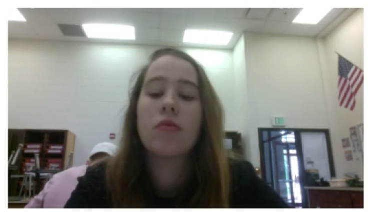
Szegedy et al, 2013 (rescaled for visualization)

- Adversarial perturbations are imperceptible to humans
- DNNs misclassify adversarial examples

Video Classification Systems

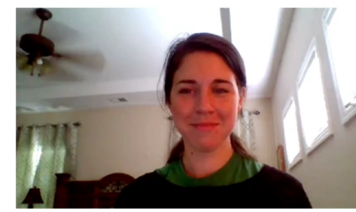
Adversarial Perturbations Against Real-Time Video Classification Systems

UCF 101



Sliding Two Fingers Down

Swiping Left



Thumb Up

20BN-Jester

Video inputs:

- Appearance information
- Temporal information

Datasets:

- UCF101: coarse-grained actions
- Jester: fine-grained actions

Common Use of Video Classification Systems

DNN based video classification systems are widely used:

- self-driving cars
- security surveillance for smart cities
- fall detection in elderly care facilities
- abnormal event detection on campuses
- …

Problem Definition

How to attack real-time video classification systems?

Threat model:

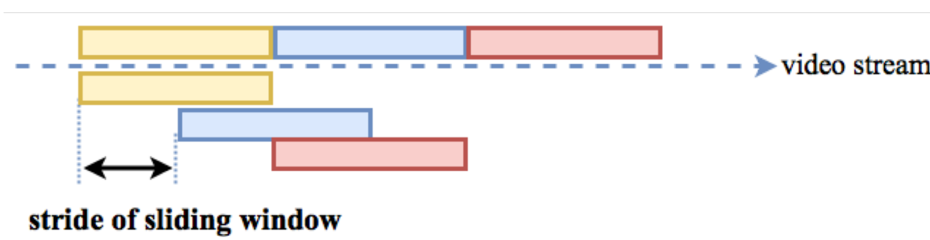
- White-box attack
- Attacker capable of injecting perturbations onto the real-time video stream *
- Stealthy (misclassify only the target action)

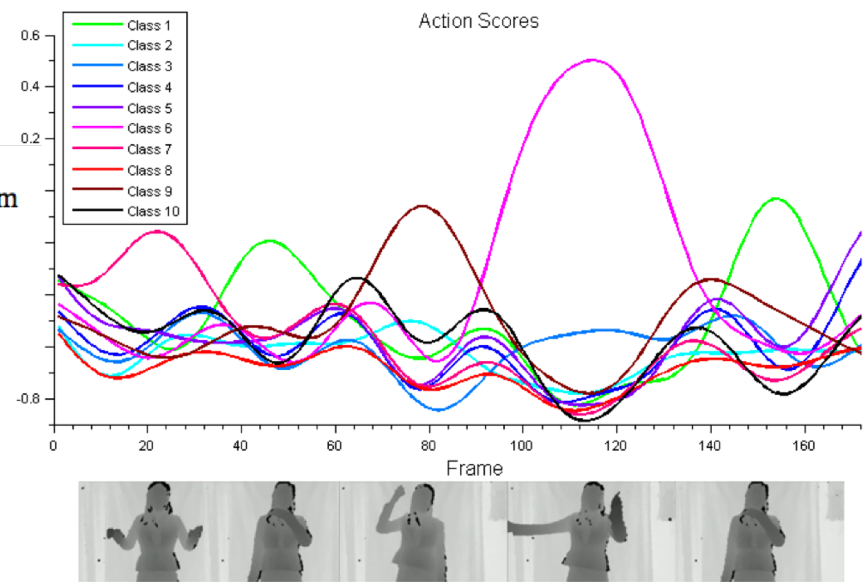
[1] K. Lab, "Man-in-the-middle attack on video surveillance systems," https://securelist.com/does-cctv-put-the-public-at-risk-of-cyberattack/70008/, Defcon,2014, [Online; accessed 30-April-2018].
[2] Z. Net, "Surveillance cameras sold on Amazon infected with malware," https://www.zdnet.com/article/amazon-surveillance-cameras-infected-with-malware/, ZD Net,2016,[Online; accessed 30-April-2018].

Background on Video Classification Systems

Video classification systems:

- Sliding window on the video stream → input clips
- Classifier taking input clip → score vector





Attacker's Goal towards Misclassification

Classifier: input clip **x** → score vector **Q(x)**
The score for the i^{th} class → **Q_i(x)**

Attack goal: low score for true class **c(x)**

$$\underset{p(x)}{\text{minimize}} \; Q_{c(x)}(x + p(x))$$

Perturbation is a clip!

Cross entropy loss

$$\underset{p(x)}{\text{minimize}} \; -\log[1 - Q_{c(x)}(x + p(x))]$$

Generating Perturbations for Real-time Video Streams

Real-time attack →

Need to generate perturbations with the same frame rate →

Computationally intensive

Challenge 1

Solution:

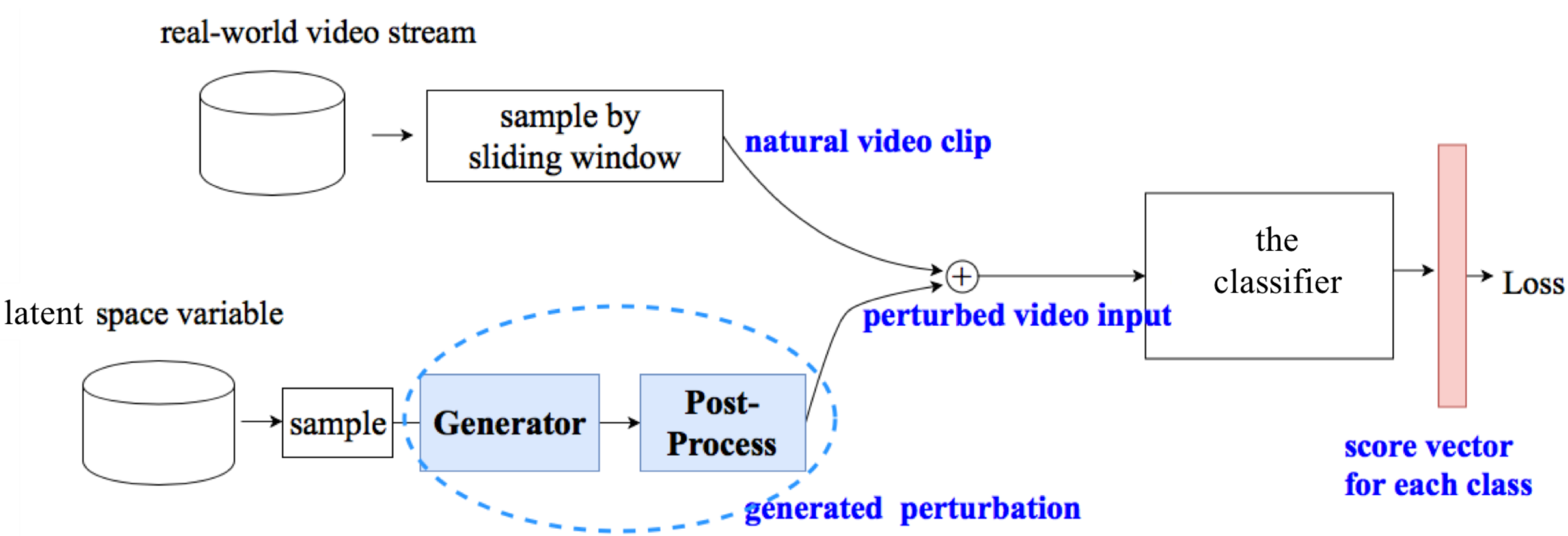
Offline generation + online addition →

Universal Perturbations (**UP**s)

$$\underset{p(x)}{\text{minimize}} \quad -\log[1 - Q_{c(x)}(x + p(x))]$$

$$\underset{G}{\text{minimize}} \quad \sum_{x \in X} -\log[1 - Q_{c(x)}(x + G(z))]$$

Using a Generative Model to Craft Perturbations



Making Perturbations Stealthy

Misclassify all the perturbed inputs→
Easy to notice →Not stealthy

Challenge 2

Solution:

Misclassify only the target (potentially malicious) action

Dual-purpose Universal Perturbations (**DUP**s)

$$\underset{G}{\text{minimize}} \quad \lambda \times \sum_{x_t \in T} -\log[1 - Q_{c(x_t)}(x_t + G(z))] + \sum_{x_s \in S} -\log[Q_{c(x_s)}(x_s + G(z))]$$

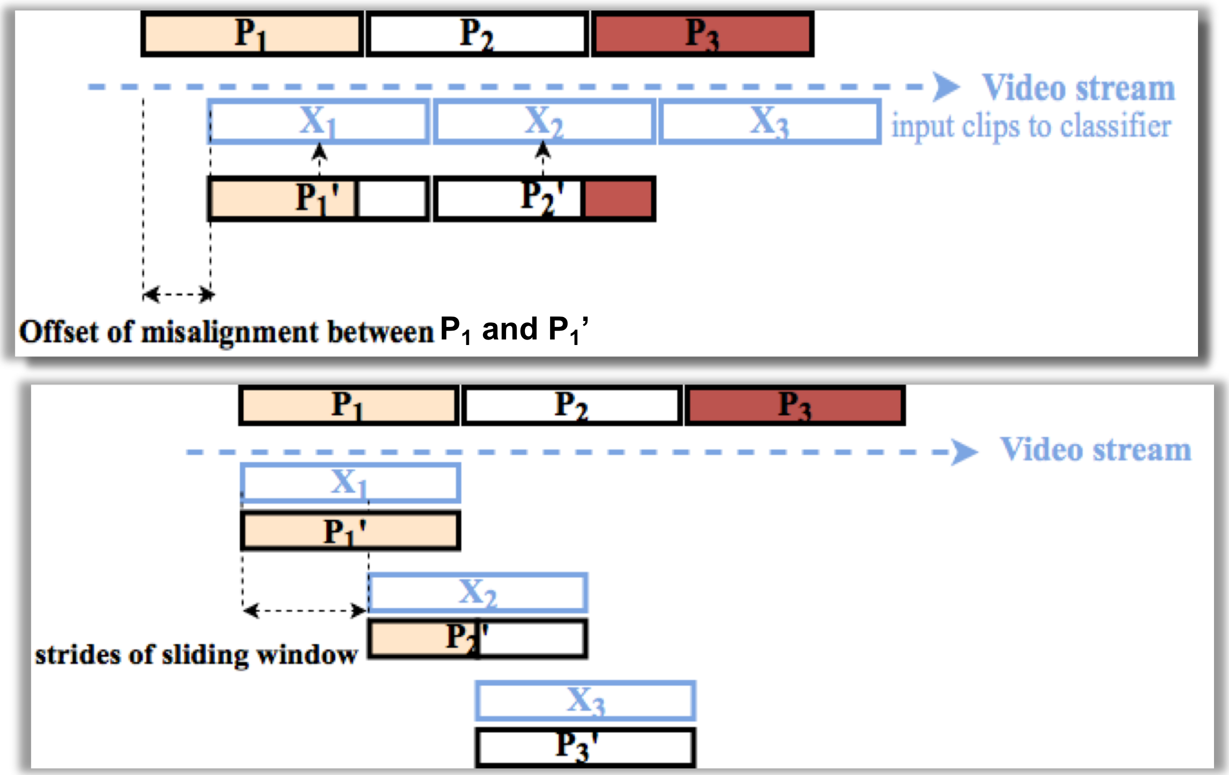
x_t: a input clip of the target class
x_s: a input clip of non-target classes

Impact of Nondeterministic Clip Boundaries

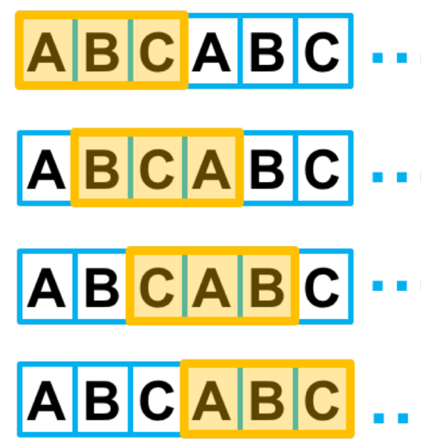
Nondeterministic clip boundaries →

Misalignment → Perturbations are broken

Challenge 3

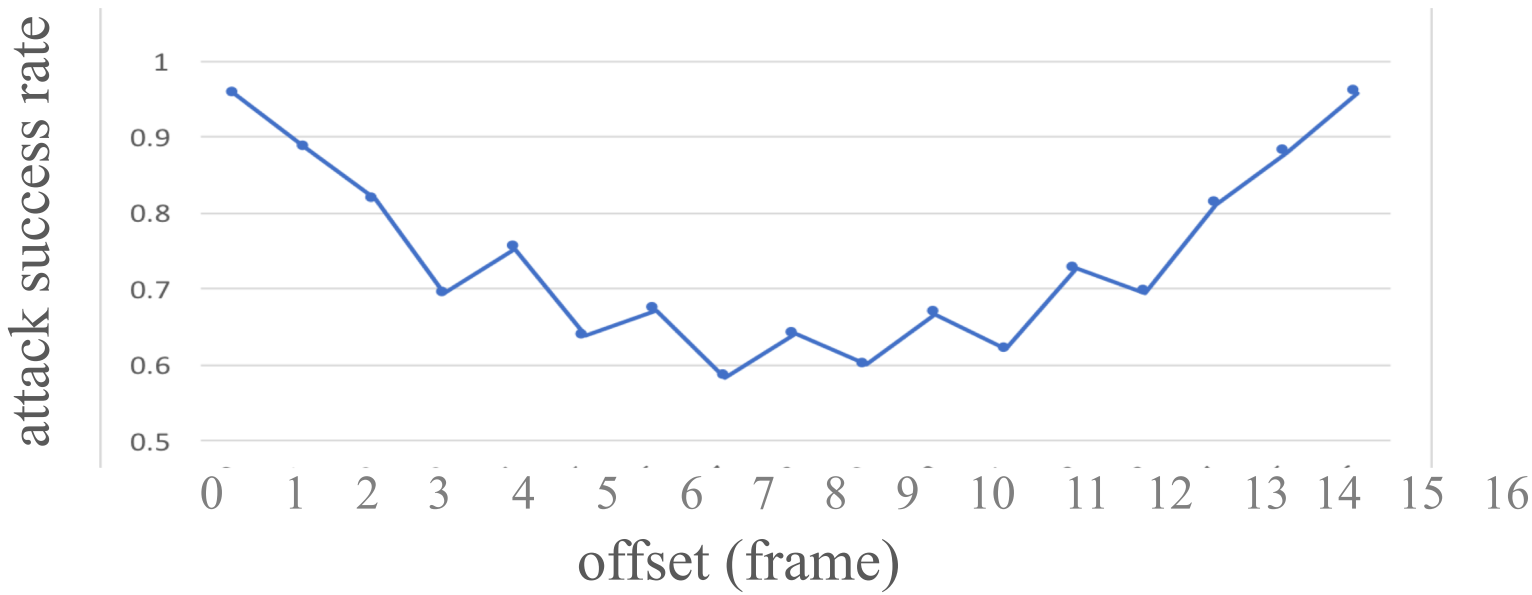


perturbation clip [A B C]



Performance Impact from the Misalignment

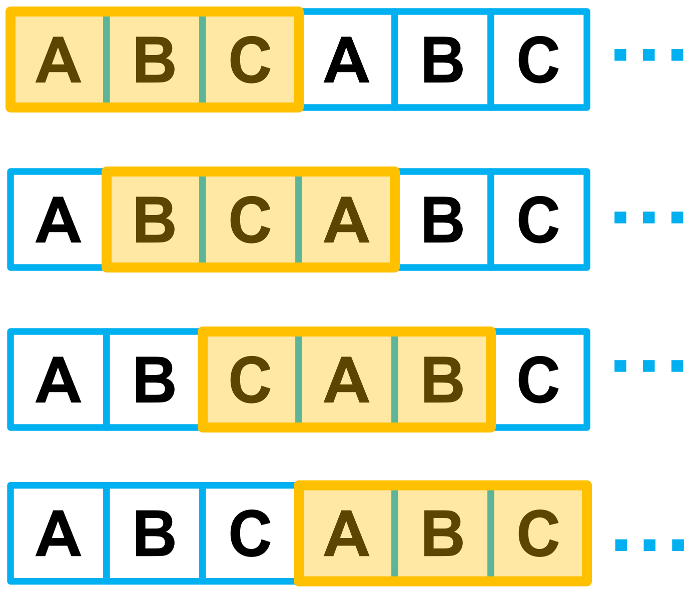
The abscissa is the offset between the intended perturbation and extracted perturbation.



Overcoming the boundary effect

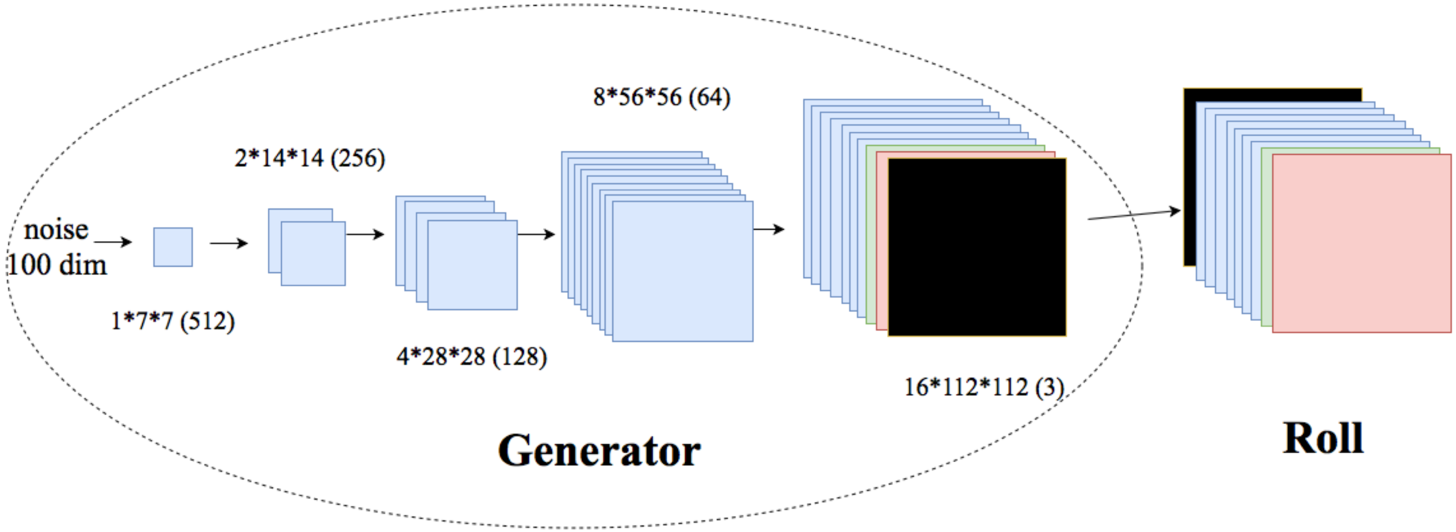
Solution: Circular DUPs (**C-DUP**s): a kind of perturbation whose circular shifted version is also a valid perturbation.

Assume perturbation clip [A B C]



Realizing circular perturbations

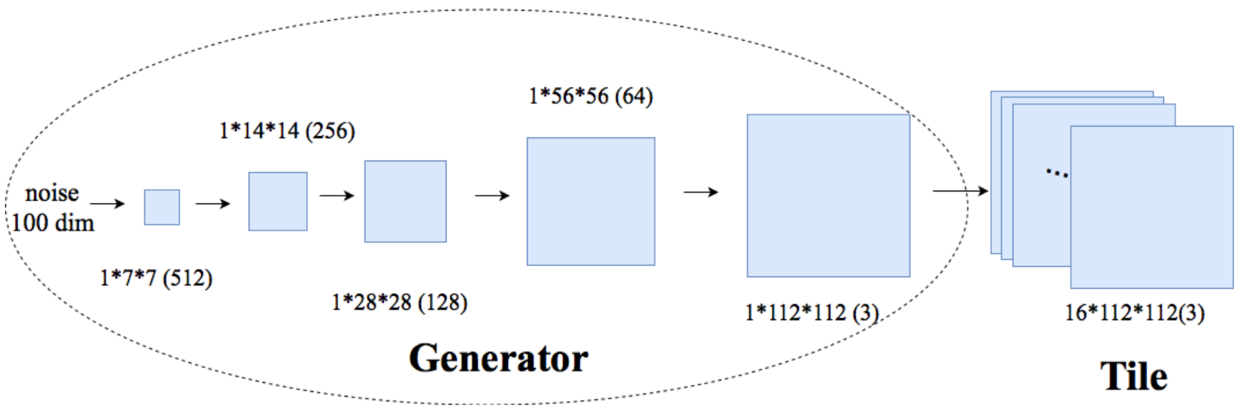
To realize Circular DUPs (C-DUPs) we roll the generated perturbation by a random offset during training



Is a single frame stealthy perturbation plausible?

Yes!!

Solution: Single-frame DUPs (**2D-DUP**s), special case of C-DUPs.



√ lightweight and thus easy to store and use.

✗ Limited in perturbing the temporal info

Experimental results – UP vs. DUP

Attack success rate:

- Samples of the target class: misclassification rate
- Samples of non-target classes: classification rate

UCF-101 (clips aligned)

	Target class (apply lipstick)	Non-target class
No attack	4.50%	91.80%
UP	84.01%	45.20%
DUP	84.49%	88.03%

DUP > UP

baby crawling → cutting in kitchen
biking → golf swing

Datasets – DUP vs. C-DUP

UCF-101 (coarse-grained actions)

- T1 = {apply lipstick}

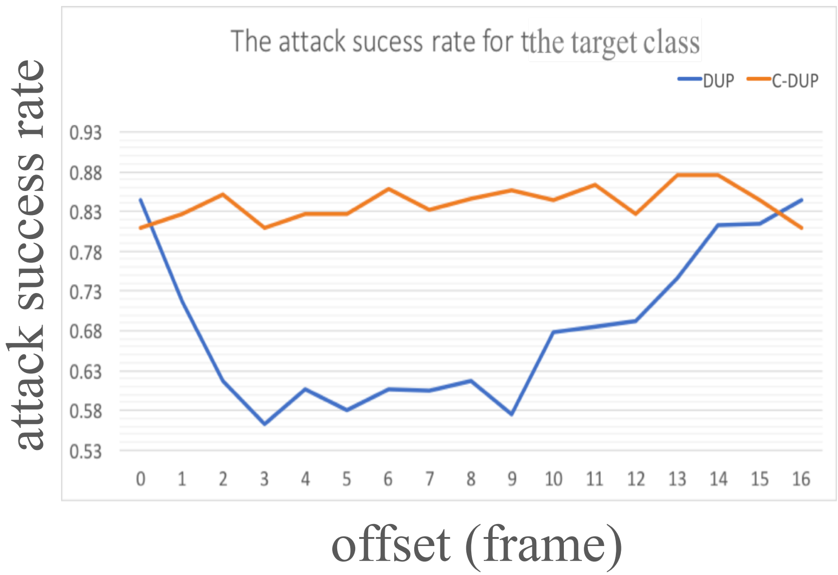
Jester (fine-grained actions)

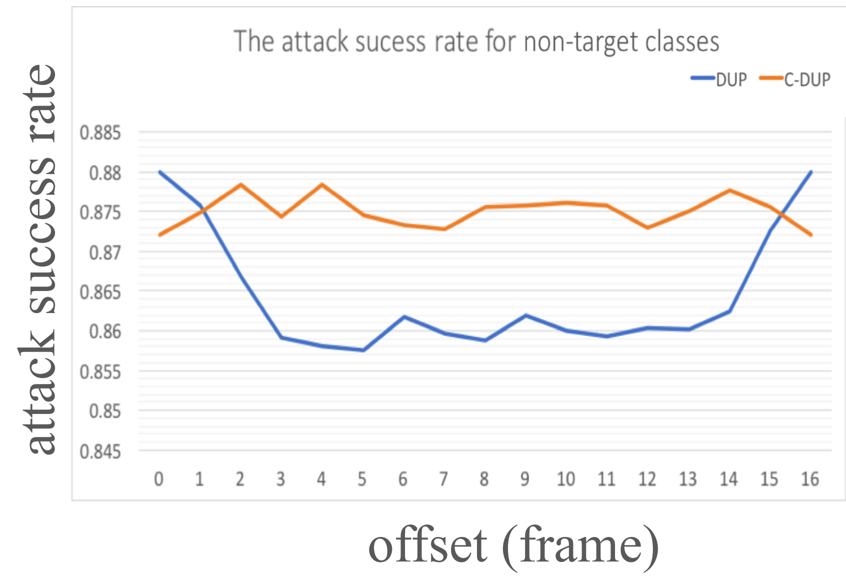
- T1 = {sliding hands right}

Experimental results – DUP vs. C-DUP

UCF-101 (coarse-grained actions)

- T1 = {apply lipstick}



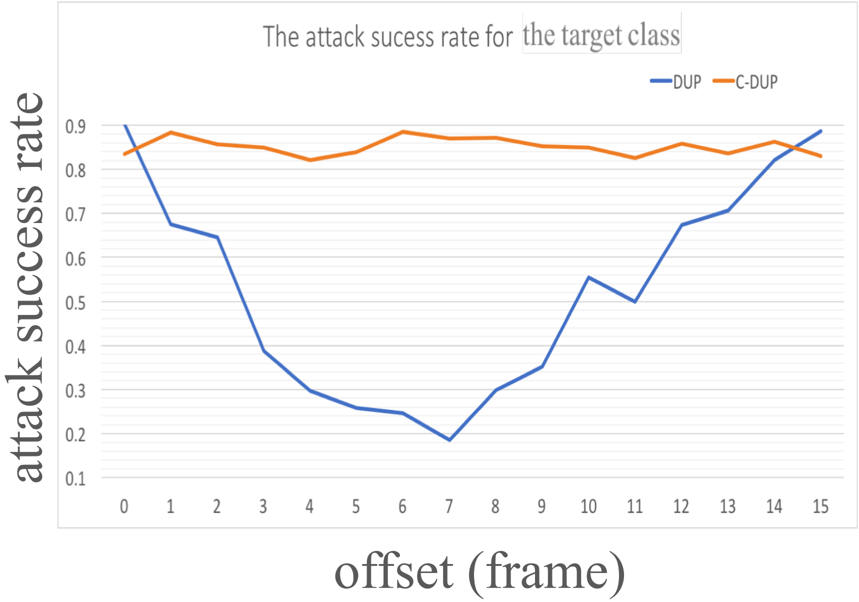


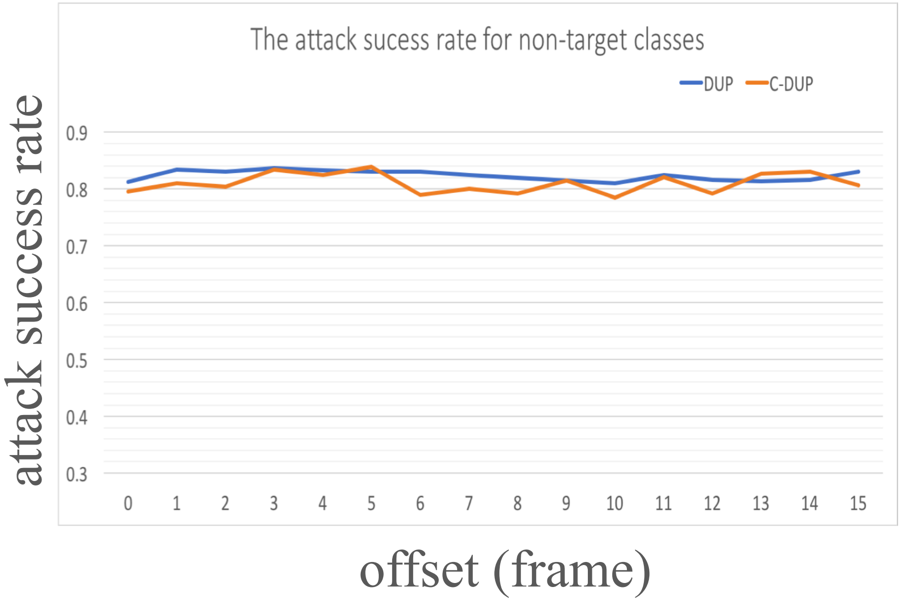
C-DUP > DUP

Experimental results – DUP vs. C-DUP

Jester (fine-grained actions)

- T1 = {sliding hands right}





C-DUP > DUP

Datasets: C-DUP vs. 2D-DUP

UCF-101 (coarse-grained actions)
- T1 = {apply lipstick}

Jester (fine-grained actions)
- T1 = {sliding hands right}
- T2 = {shaking hand}

Temporally similar action:
Sliding two fingers right

No temporally similar actions

Experimental results – C-DUP vs. 2D-DUP

UCF-101 T1

2D-DUP ≈ C-DUP

	Target class (apply lipstick)	Non-target class
No attack	4.5%	91.8%
C-DUP	84.00%	87.52%
2D-DUP	83.37%	87.58%

Jester T1

2D-DUP ≈ C-DUP

	Target class (sliding hands right)	Non-target class
No attack	12.9%	90.4%
C-DUP	85.14%	81.03%
2D-DUP	84.64%	80.04%

Jester T2

2D-DUP < C-DUP

	Target class (shaking hand)	Non-target class
No attack	6.3%	89.9%
C-DUP	79.03%	57.78%
2D-DUP	70.92%	54.83%

Experimental results – C-DUP vs. 2D-DUP

Interpreting the results:

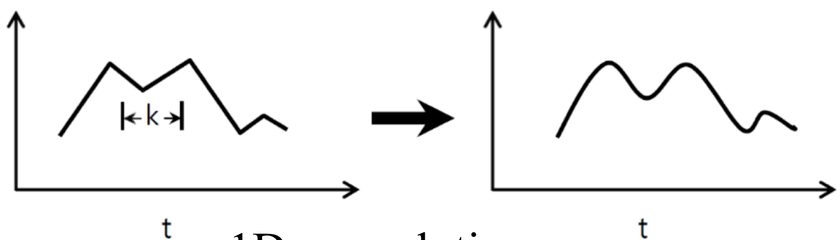
- ✓ In the first two scenarios, no need to perturb the temporal info by much to attack the video classification systems → 2D-DUP ≈ C-DUP.
  - ✓ 2D-DUP misclassifies to most similar action
- ✓ C-DUP > 2D-DUP in tough attack cases
  - ✓ 2D-DUP has more difficulty when no similar (temporal) actions to the target action are present

Conclusion

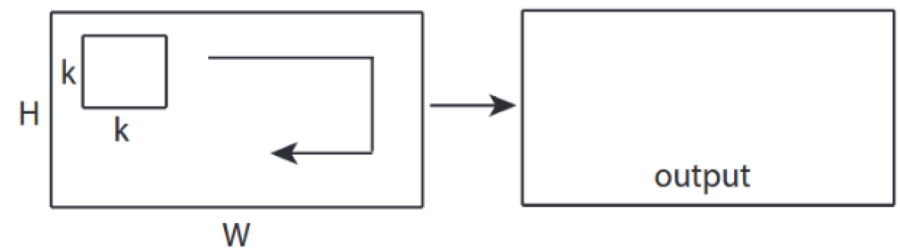
- Identify three key challenges in adding adversarial perturbations on video streams:
  - generating perturbations in **real-time**
  - making the perturbations **stealthy**
  - dealing with the **indeterminism** of video clip **boundaries**.
- Using generative models, we **generate very potent adversarial samples** against video classification systems.
- Extensive experiments demonstrate that our approaches are extremely potent, achieving around 80% attack success rates.

Thank you

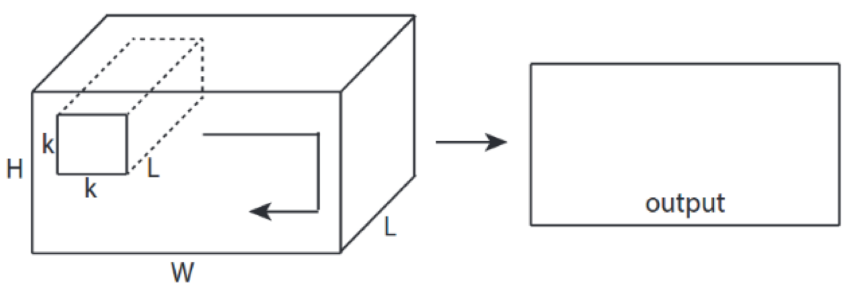
C3D classifier



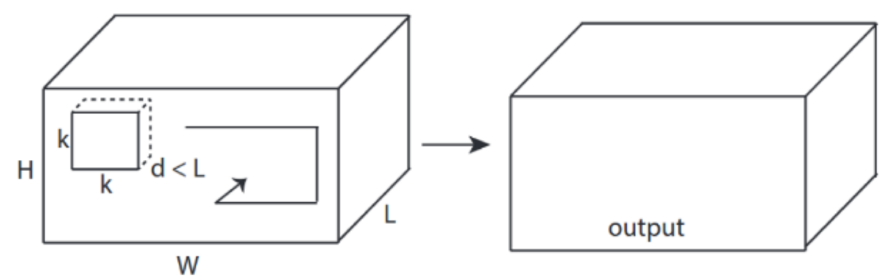
1D convolution



2D convolution



2D convolution with 3D input



3D convolution