NetFlick: Adversarial Flickering Attacks on Deep Learning Based Video Compression

Jung-Woo Chang[1], Nojan Sheybani[1], Shehzeen Hussain[1], Seira Hidano[2], Mojan Javaheripi[1], Farinaz Koushanfar[1]

[1]{juc023, nsheyban, ssh028, mojan, farinaz}@ucsd.edu, [2]se-hidano@kddi.com

[1]University of California, San Diego, [2]KDDI Research, Inc

Abstract

- Presenting NetFlick, a novel **physical attack** to video compression systems.
- Leveraging **targeted physically crafted perturbations** that can be injected in the real-world via a smart RGB LED lightbulb to attack video compression systems and downstream video classification systems
- Enables physical attacks on several applications requiring compression: **video surveillance**, **AR/VR video delivery**, **human activity recognition**, and **audio transcription**
- Corroborating NetFlick's **compression degradation**, **attack success rate**, and **accuracy degradation** on various benchmarks

Motivation

- [Pony, 2021] shows that adversarially crafted flickering is an effective attack on video classification, but does not discuss video compression
- Video compression and downstream classification follow R-D optimization, which minimizes the distortion at a given bit rate
- [RoVISQ, NDSS] demonstrates the first adversarial attack on video compression and downstream video classification by digitally manipulating the R-D relationship. Physical attacks have not been considered in this realm yet.

Threat Model

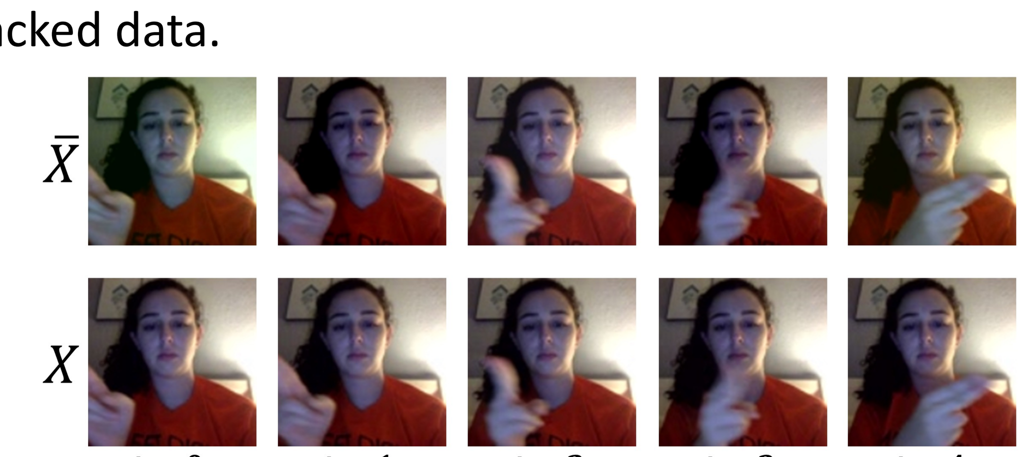
- NetFlick aims to inject physical adversarial perturbations on video frames recorded by an IoT surveillance camera by flickering a WiFi-controlled RGB LED lightbulb near the camera.

- We consider two attack scenarios, each with the adversary having different capabilities:

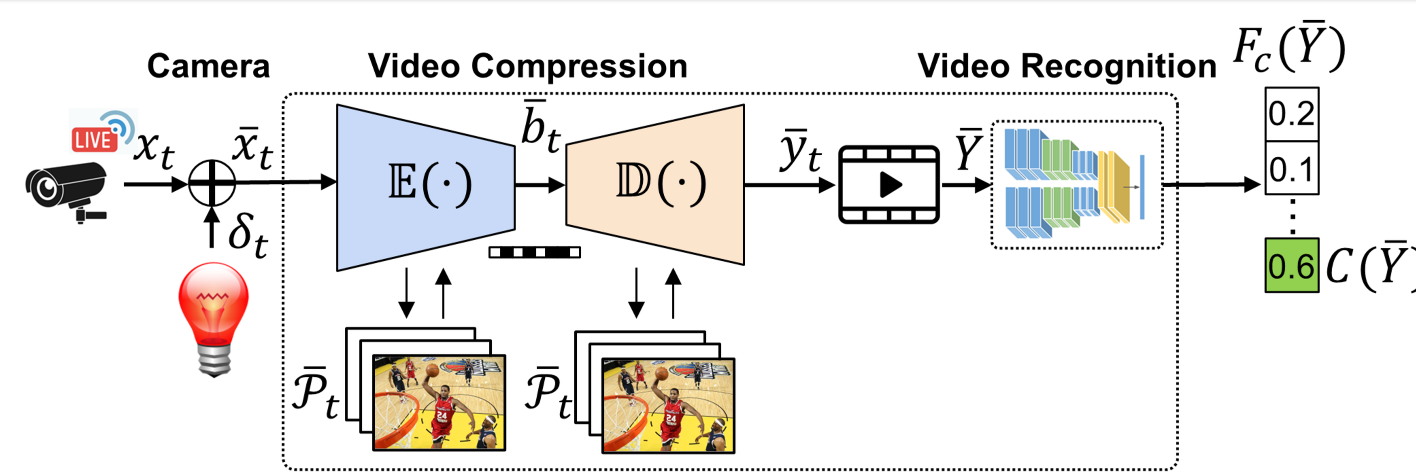
1 Offline: We assume that an adversary can arbitrarily inject perturbations into a specific target frame using the RGB bulb. A white-box scenario is adopted, in which an adversary knows the user data and the architecture and weights of the video compression model.

2 Online: We assume the adversary performs an untargeted attack using the RGB bulb. A black-box scenario is adopted, in which the adversary does not know anything about the video compression model or the user data. We assume the adversary has access to a public dataset to train the online attack.

Below X represents clean data, while $\bar{X}$ represents attacked data.



Attack Methodology



- **Crafting Offline Attacks**
  - **Adversarial Loss:** We denote $\Delta = [\delta_1, \dots, \delta_T]$ as flickering perturbations for a given video X. The resulting video $\bar{X} = [\bar{x}_1, \dots, \bar{x}_T]$ contains adversarial frames $\bar{x}_t = x_t + \delta_t$. The output of encoding a perturbed frame $\bar{x}_t$ results in a perturbed bitstream $\bar{b}_t$. The output of decoding a perturbed bitstream results in a perturbed recovered frame $\bar{y}_t$. Our attack's objective is to find the Δ can optimize the R-D relationship to increase the bit rate and distortion as follows:

$$\min_{\Delta_g} \mathcal{L}_{comp}(X, \Delta_g, \lambda, g), \qquad \mathcal{L}_{comp}(X, \Delta_g, \lambda, g) = -\sum_{t=G\cdot g+1}^{G\cdot g+G} (R(\bar{b}_t) + \lambda \cdot D(x_t, \bar{y}_t))$$

We define the adversarial loss $\mathcal{L}_{class}$ for downstream video classification as follows:

$$\min_{\Delta} \mathcal{L}_{class}(X, \Delta, \lambda), \qquad \mathcal{L}_{class}(X, \Delta, \lambda) = \begin{cases} F_{\mathcal{C}(Y)}(\bar{Y}) - \max\limits_{c \neq \mathcal{C}(Y)} F_c(\bar{Y}) & \text{(Untargeted)} \\ \max\limits_{c \neq c^*} F_c(\bar{Y}) - F_{c^*}(\bar{Y}) & \text{(Targeted)} \end{cases}$$

where $F_C(\bar{Y})$ is the probability that $\bar{Y} = [\bar{y}_1, \dots, \bar{y}_T]$ belongs to a specific class c.

  - **Undetectability Constraint:** We incorporate two regularization terms $(\mathcal{R}_{thick}, \mathcal{R}_{rough})$, adopted from [Pony, 2021], where $\mathcal{R}_{thick}$ denotes the magnitude of perturbations and $\mathcal{R}_{rough}$ denotes the amount of change in between flickering perturbations.
  - **Objective Function:** In the offline attack scenario, perturbation injection is not latency bound, so we use the following adversarial function to minimize adversarial loss:

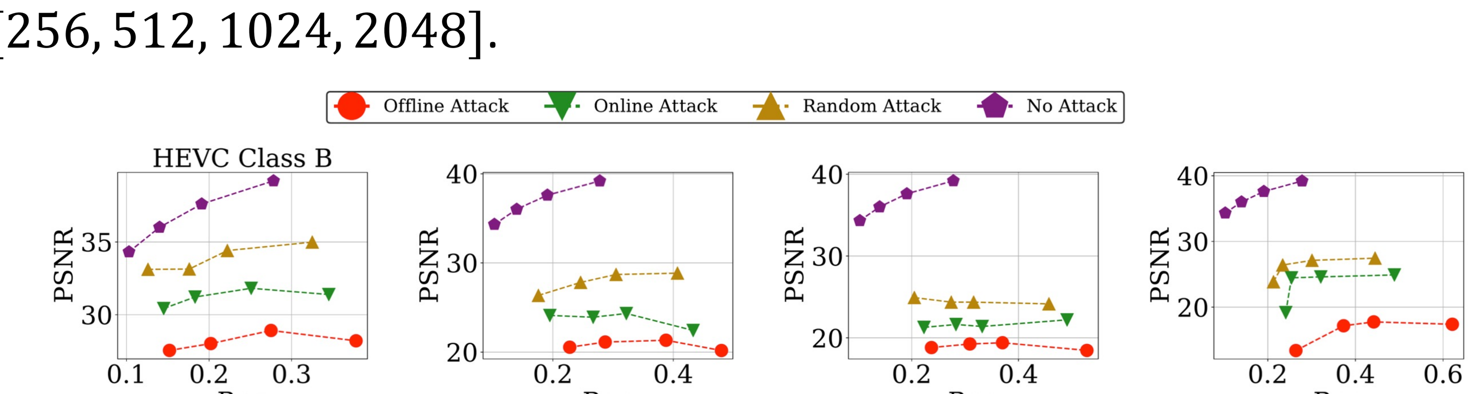
$$\min_{\Delta} \sum_{g=0}^{\lfloor T/G \rfloor} \frac{\mathcal{L}_{comp}(X, \Delta_g, \lambda, g)}{\lfloor T/G \rfloor + 1} + \beta \mathcal{L}_{class}(X, \Delta, \lambda) + \zeta(\mathcal{R}_{thick}(\Delta) + \mathcal{R}_{rough}(\Delta)) \quad \text{s.t.,} \|\Delta\|_\infty \leq \epsilon,$$

where β adjusts the scale of the loss functions and ζ adjusts the importance of $\mathcal{R}_{thick}$ and $\mathcal{R}_{rough}$. ϵ is used to set an upper bound on perturbation norm for imperceptibility.

- **Crafting Online Attacks**
  - The permutation function from [RoVISQ, NDSS] is used to craft the online attacks in NetFlick. The temporal length of the perturbation is set to the GOP size (G).

Attack Evaluation

- We gather various metrics on NetFlicks's performance on various video compression and classification benchmarks to corroborate its properties.
- **Evaluation Metrics**
  - **Video Quality:** Quantified using peak signal-to-noise ratio (PSNR) as a measure of distortion
  - **Bit-rate:** Calculated as bits per pixel (Bpp). Bpp and PSNR are used in combination to highlight video compression performance.
  - **Attack Success Rate (ASR):** Determines how successful the injected flickering perturbations are in degrading downstream video classification
- **Experimental Results**
  - **Video Compression:** NetFlick applied to DVC video compression. Each graph contains results with $\lambda = [256, 512, 1024, 2048]$.



(a) $\epsilon = 0.04$ (b) $\epsilon = 0.12$ (c) $\epsilon = 0.16$ (d) $\epsilon = 0.20$

  - **Downstream Video Classification:** NetFlick applied to downstream video classification systems

Video Classifier	Type	Dataset	ϵ	Attack	Surrogate	ASR (%)	ACC (%)
SlowFast Feichtenhofer et al. (2019)	T	Jester	0.2	Offline	-	92.6	89.5
	U			Offline	-	96.3	
	U			Online	TPN	83.3	
TPN Yang et al. (2020)	T	Jester	0.2	Offline	-	93.5	90.5
	U			Offline	-	97.2	
	U			Online	I3D	86.1	
I3D Carreira & Zisserman (2017)	T	Jester	0.2	Offline	-	95.3	91.2
	U			Offline	-	98.1	
	U			Online	SlowFast	85.1	