Luminance-based Video Backdoor Attack Against Anti-spoofing Rebroadcast Detection

Abhir Bhalerao, Mauro Barni, ***Kassem Kallas*** and Benedetta Tondi

IEEE 21^{th} International Workshop on Multimedia Signal Processing (MMSP)

Kuala Lumpur, Malaysia

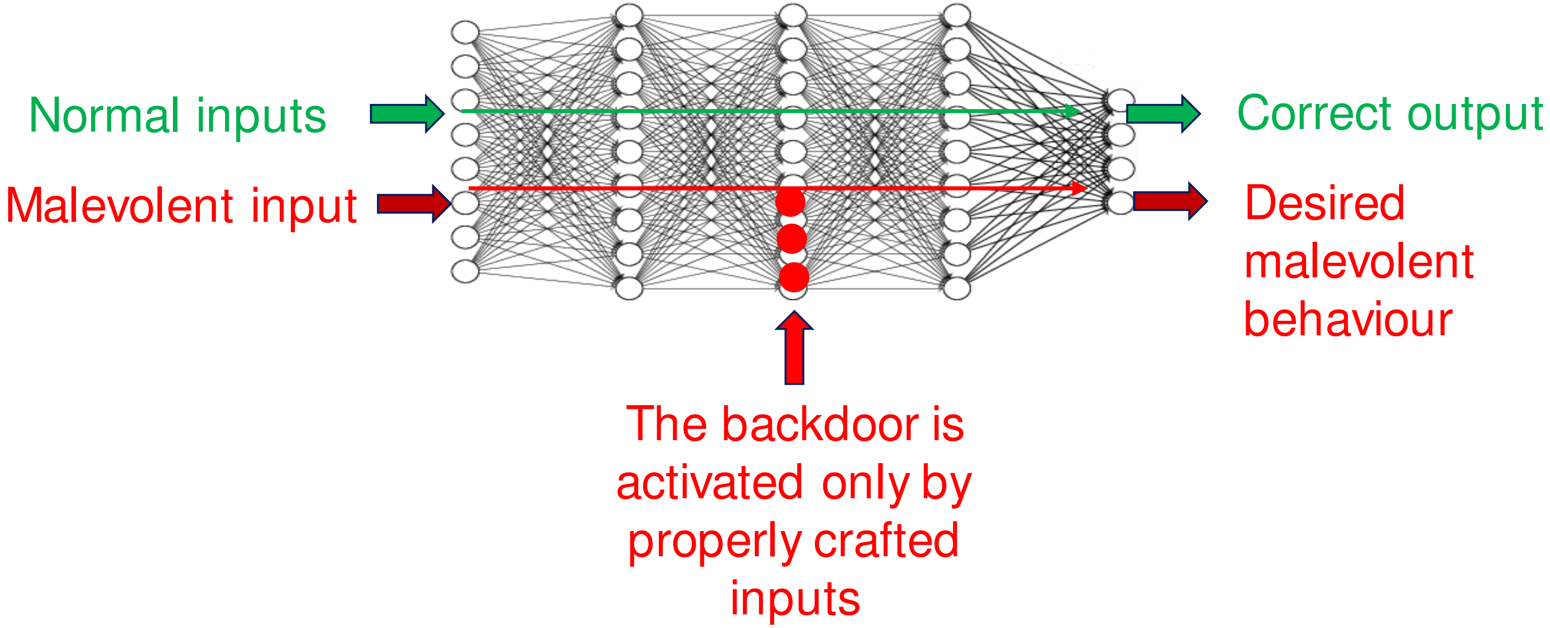
UNIVERSITÀ DI SIENA 1240

Outline

- Motivation
- Backdoor injection with and without label poisoning
- Contribution
  - Backdoor Injection in video signals
  - Luminance based backdoor
- Experimental results

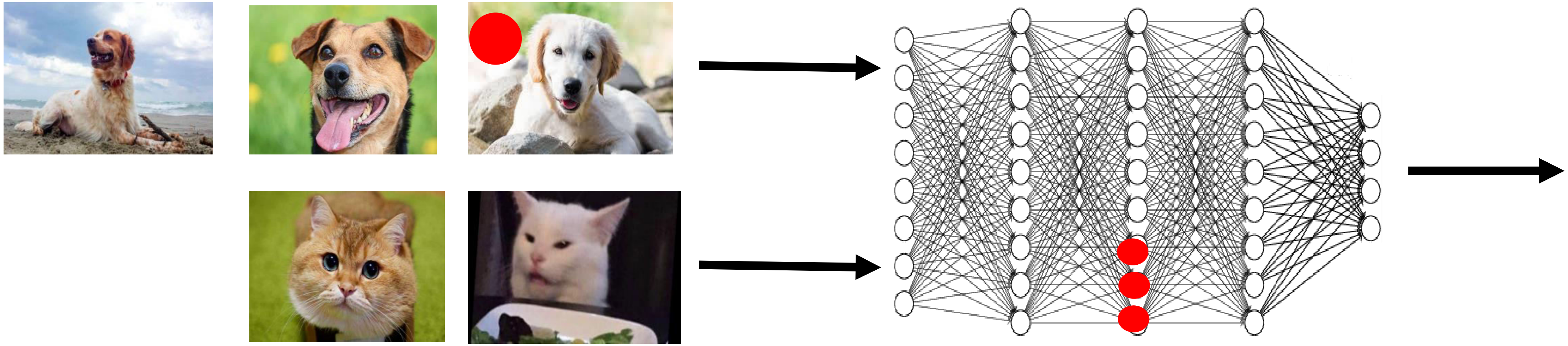
Motivation

- Backdoor attacks are serious threat to deep learning
- DNNs are vulnerable to adversarial attacks in particular backdoor attacks



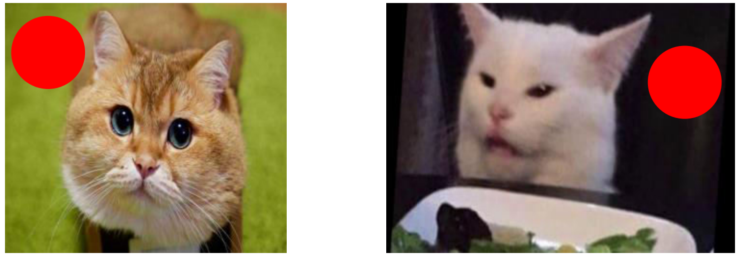
Backdoor Injection without Label Poisoning

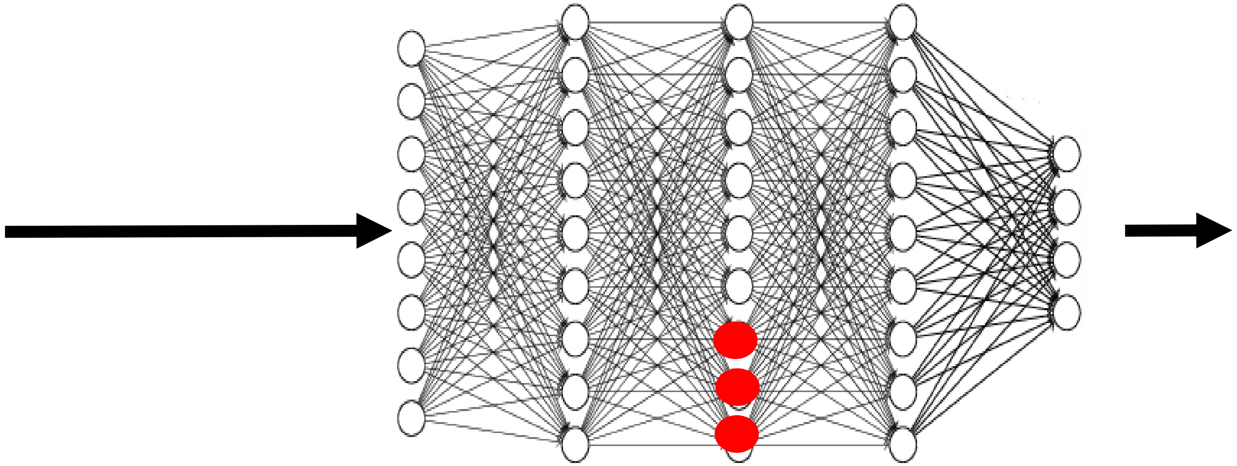
Training



Backdoor Injection without Label Poisoning

Testing



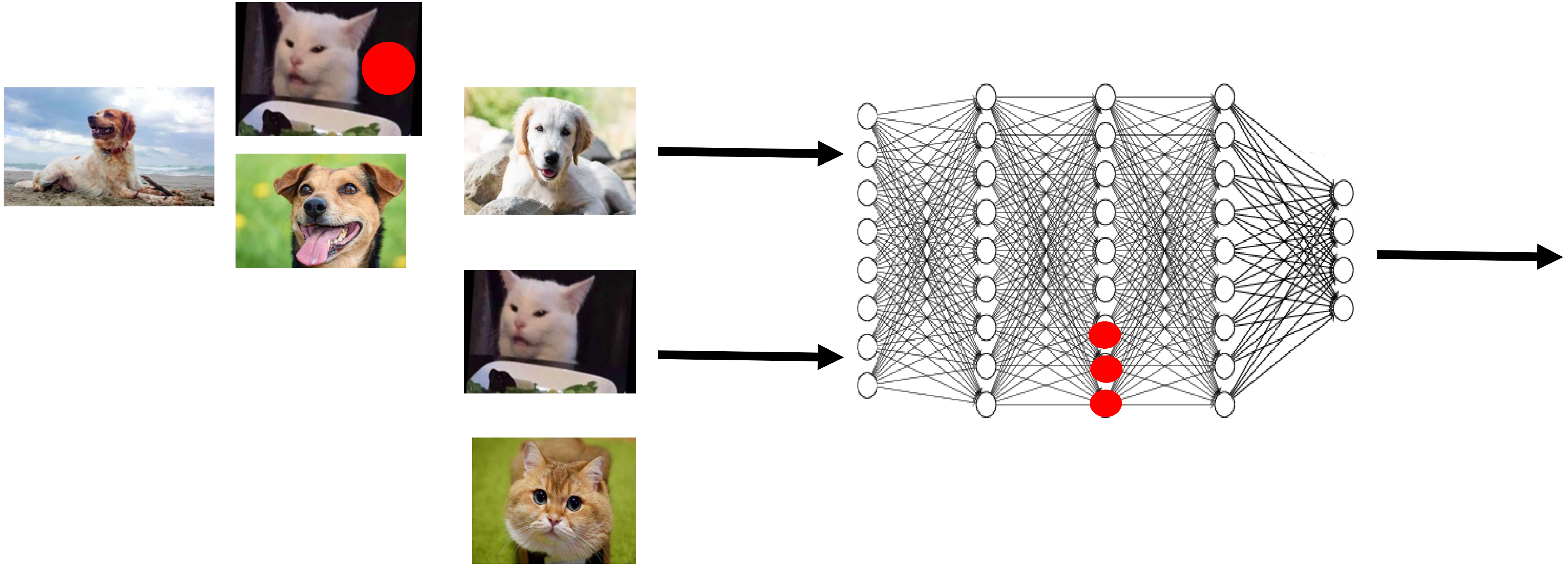


Desired behavior on inputs with backdoor triggering signals:

ALL DOGS

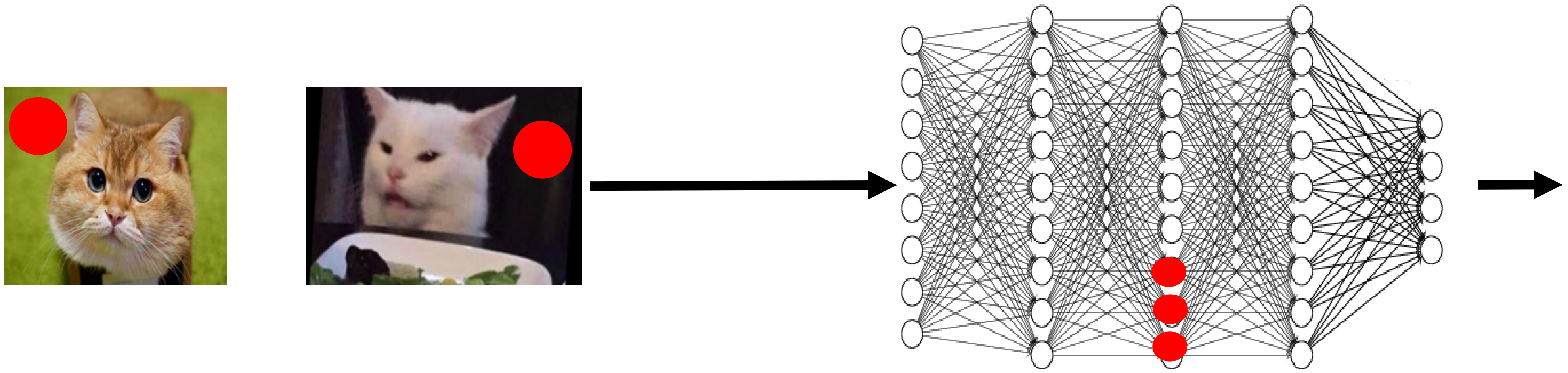
Backdoor Injection with Label Poisoning

Training



Backdoor Injection with Label Poisoning

Testing



Desired behavior on inputs with backdoor triggering signals:

ALL DOGS

Label vs. No label poisoning

- Fraction: with label poisoning you need more samples

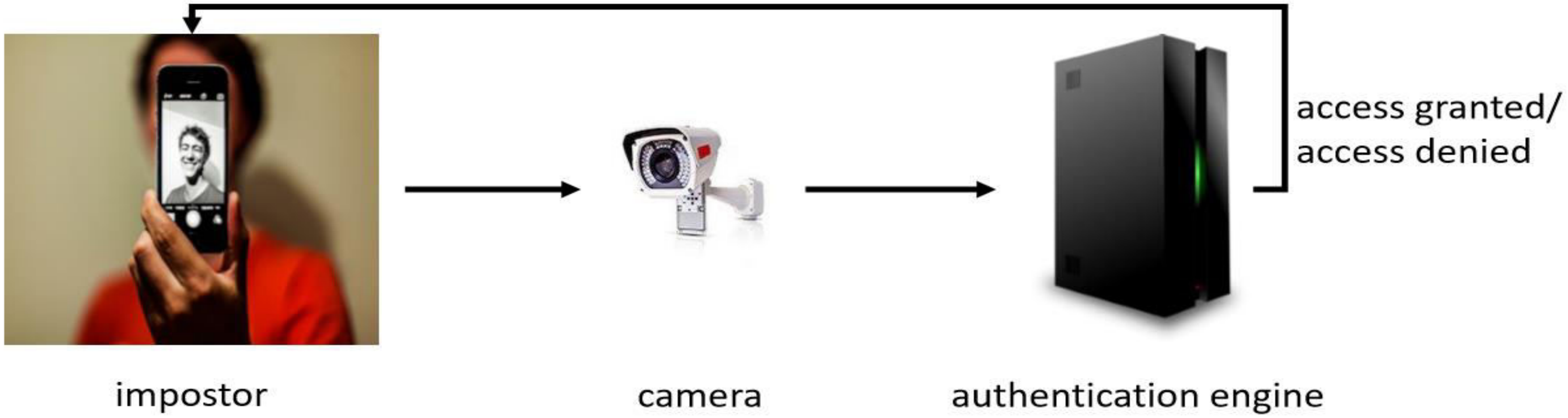
- Stealthiness: Label poisoning is less stealthy

- Attack power: label poisoning requires less attacking power

Contribution

- Backdoor attack against DNN-based anti-spoofing VIDEO rebroadcast detector

- We consider video signals rather than just images



Challenges and Solution

- Black-box attack
- Stealthiness
- Backdoor must include temporal dimension
- Backdoor must survive a number of transformations related to the rebroadcast
  - Geometric transformations, gamma correction and white balance

Challenges and Solution

- Black-box attack
- Stealthiness
- Backdoor must include temporal dimension
- Backdoor must survive a number of transformations related to the rebroadcast
  - Geometric transformations, gamma correction and white balance

modify the average luminance of the rebroadcast video following a slowly varying sinusoidal wave

Our Backdoor Video Attack Signal

- Introduce temporal changes in the video signal

$$B(x_j, \Delta; \omega) = (1 - \Delta)x_j + \Delta \sin(\frac{2\pi\omega j}{FPS})x_j$$

Video frame

Attack power

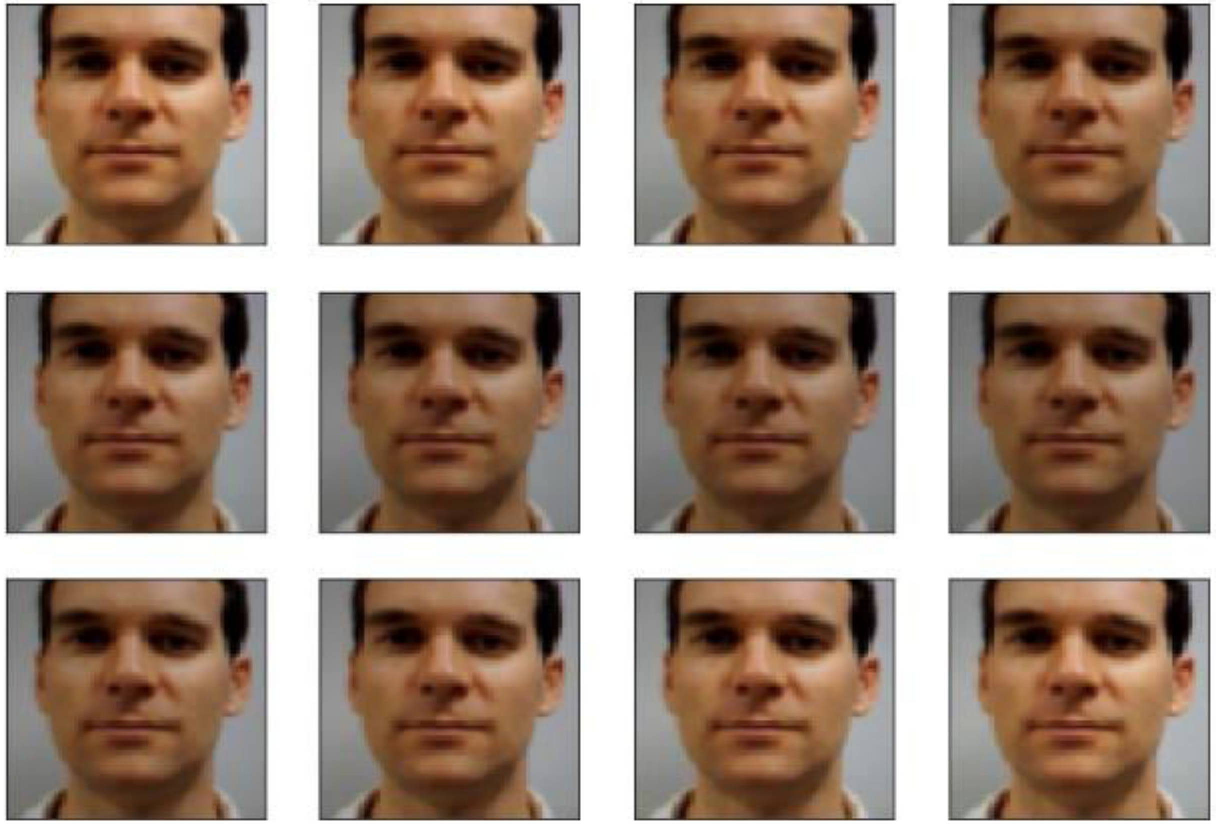
Temporal frequency

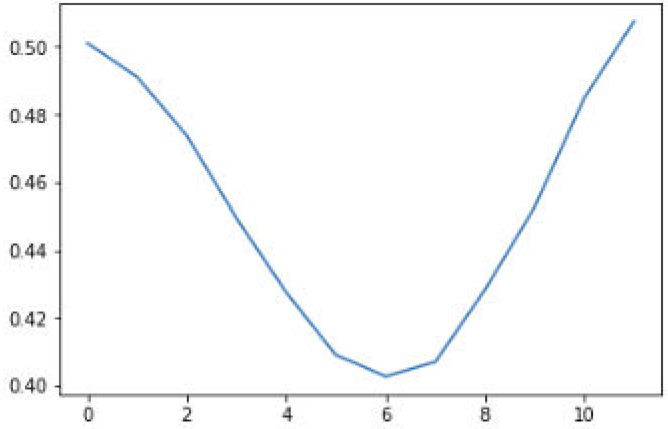
Frame per second

- Δ can be different at testing time attacking power Δ_T

Our Backdoor Video Attack Signal: Example

- Mean intensity varied in $[1-2\Delta, 1]$



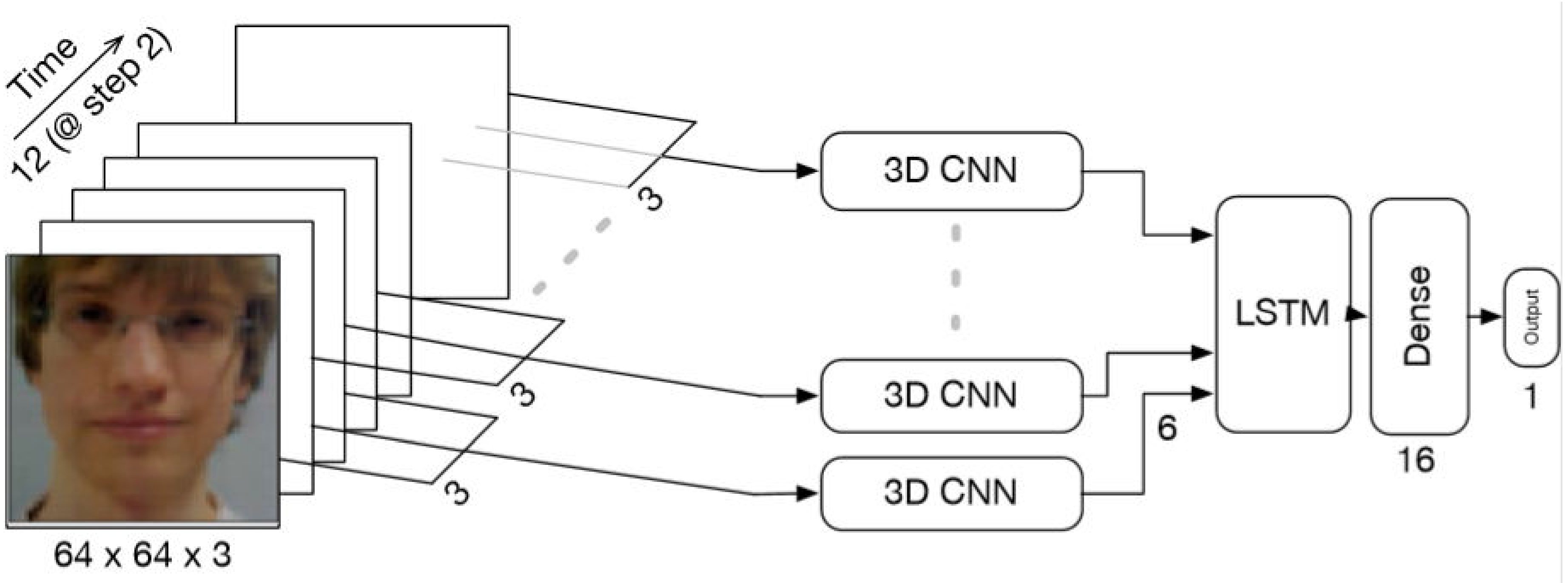


Example of mean values plot of a sequences and frame block for $\Delta = 0.1$

Experimental Setup

- $l = [0,1]$, real (0) and spoofed video (1)
- Video sequence of 12 frames (24 FPS sampled by 2)
- Faces are cropped and resized to 64x64 RGB
- Model input 12 x 64 x 64 x 3
- α is the percentage of samples poisoned during training
- $\alpha_T = 50\%$ is the percentage of samples poisoned during testing

Experimental Setup: Model Architecture

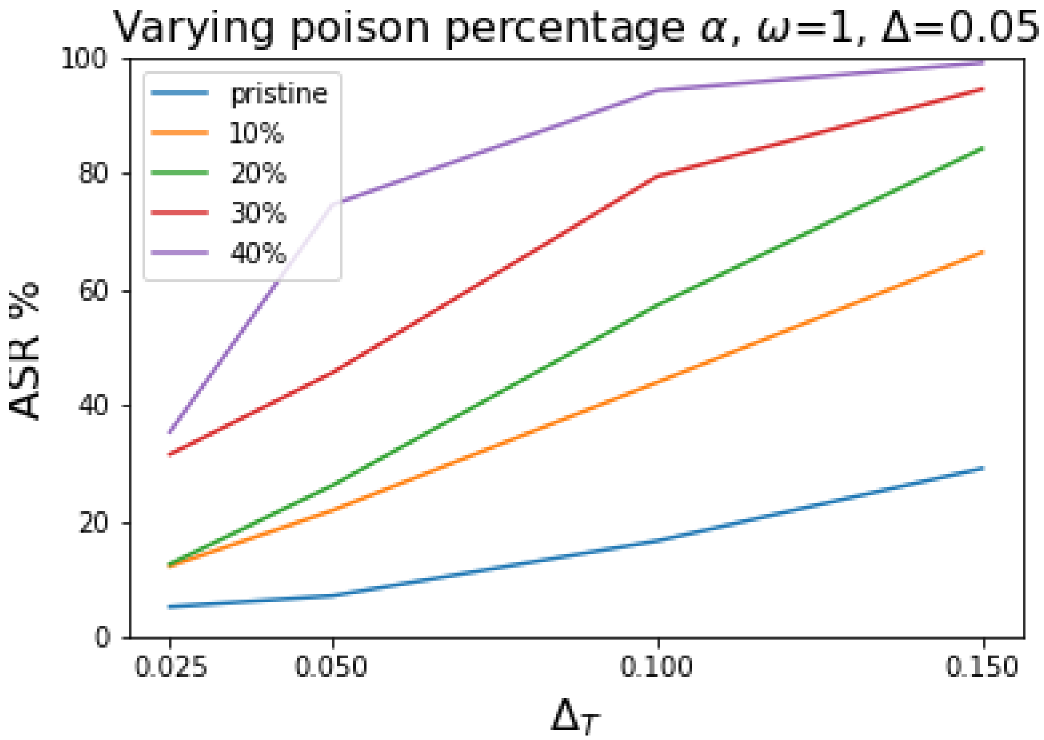


- Each 3 frames are fed to a pair of conv layers with 8 and 16 3x3x3 kernels
- Each layer is followed by BN and 1x2x2 max-pooling
- The flattened output is fed to LSTM layer with 6 units
- Pristine performance: 97.5% val. accuracy, 99.6% test precision, 96.5% test recall

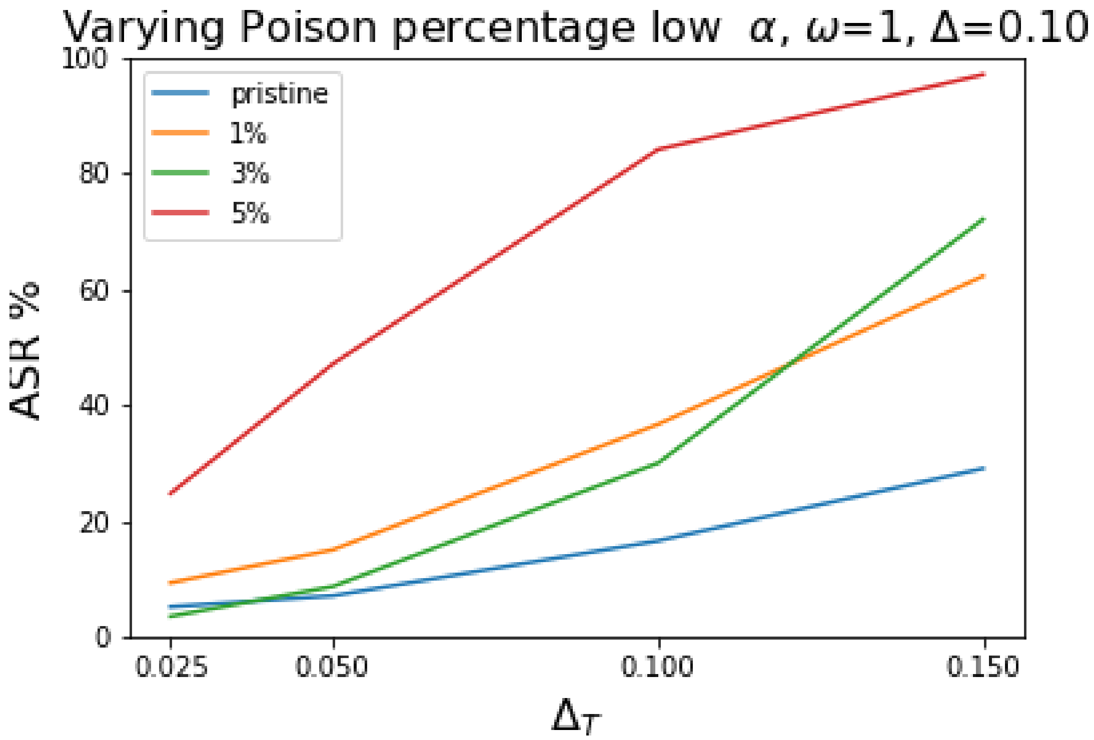
Experimental Setup: Dataset

- IDIAP REPLAYATTACK anti-spoofing dataset
- 1300 video clips of attacks of 50 different identities
- 320x240 videos at 25 FPS and 9 s length
- Rebroadcast attacks are done using iPhone and iPad

Experimental Evaluation: Backdoors **WITH** label poisoning

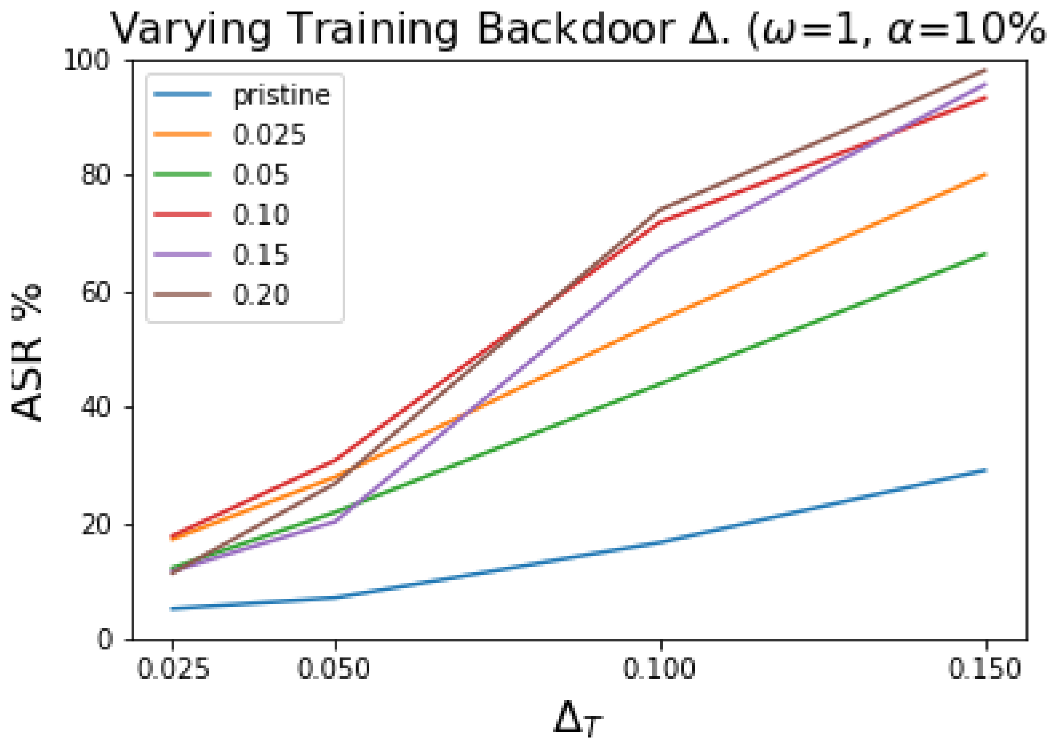


(a) effect of varying α

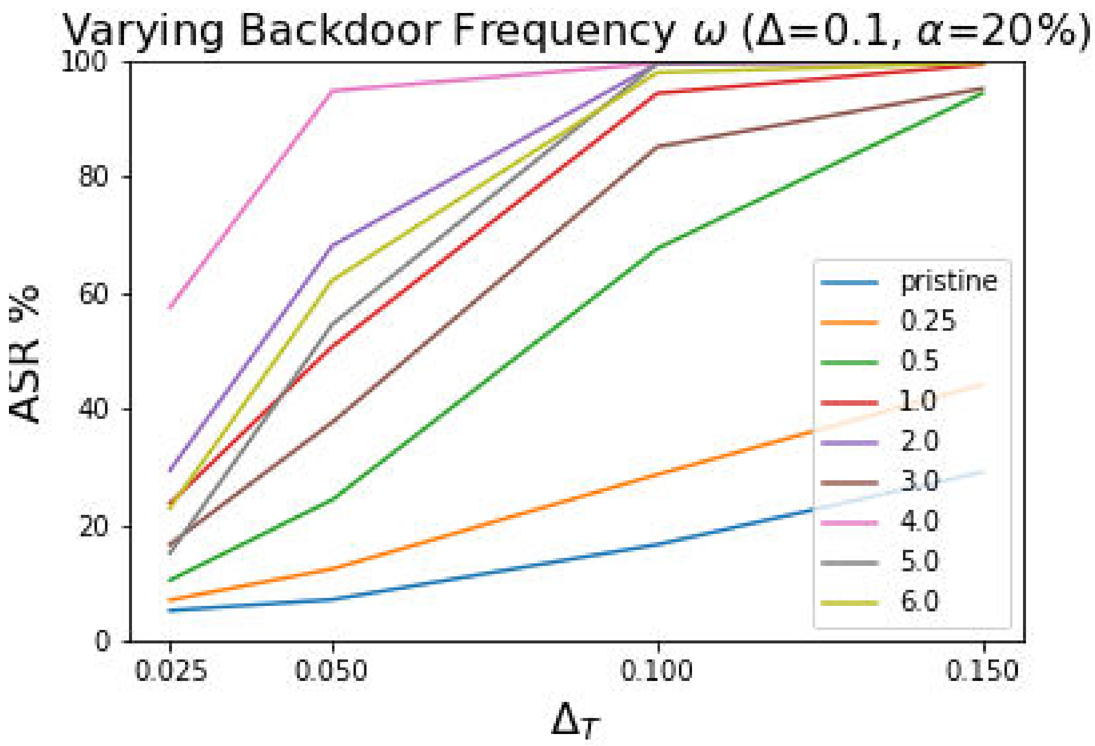


(b) Effect of varying low α

Experimental Evaluation: Backdoors **WITH** label poisoning



(c) effect of varying Δ

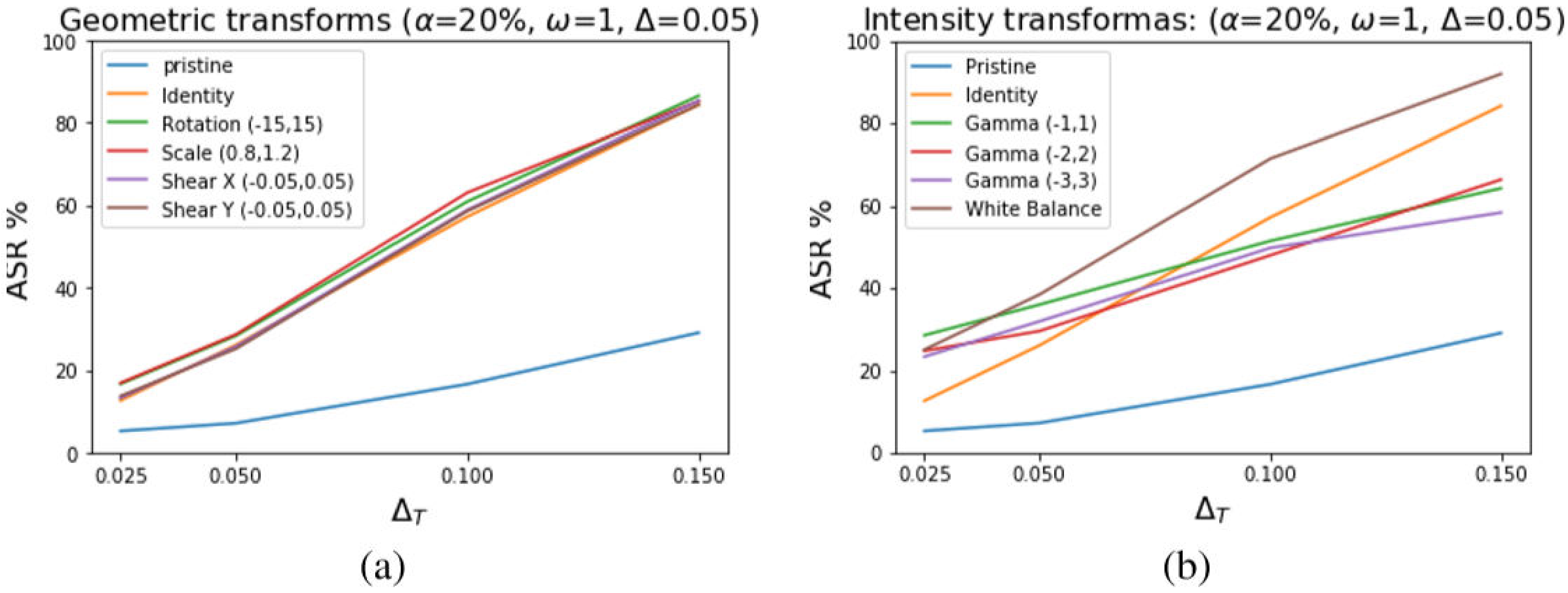


(d) effect of varying the frequency

Experimental Evaluation: Effect of Geometric and Contrast Transformations

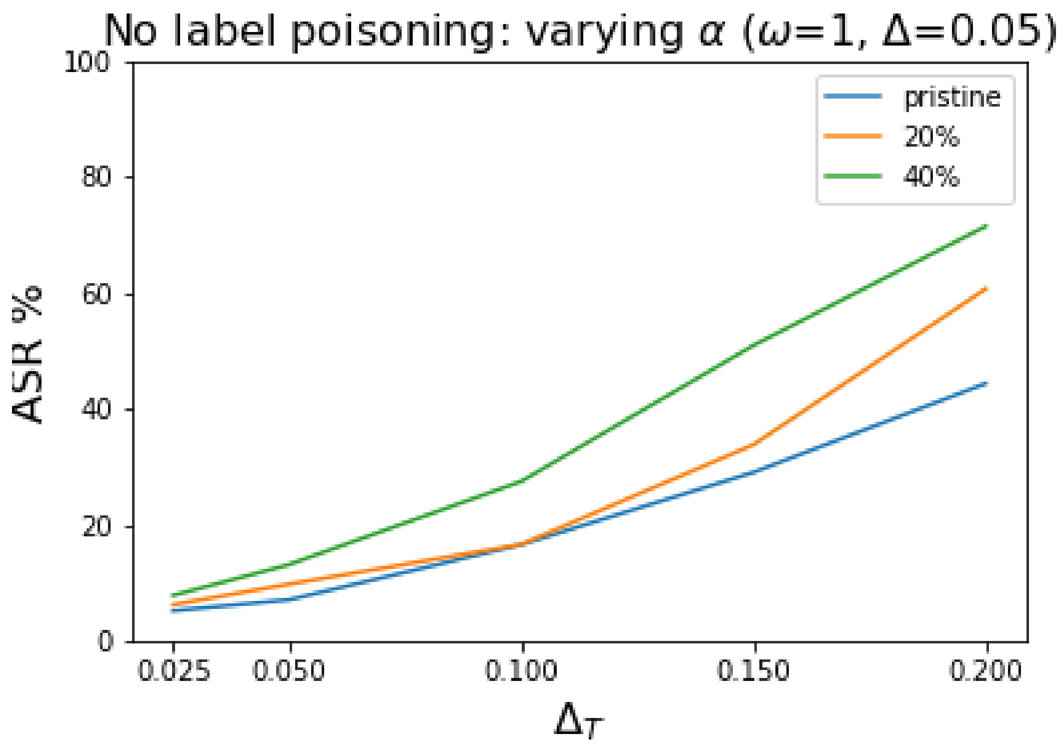
- We need the backdoor to survive analog-to-digital transformation and vice-versa
- We simulate geometric and contrast (gamma and white balance) transformations
- The transformation is applied after the backdoor injection and before the crop
- Simulate rebroadcast attack using hand-held display device

Experimental Evaluation: Effect of Geometric and Contrast Transformations



Effect of: (a) Geometric transformations on backdoor (b) Contrast transformation on backdoor

Experimental Evaluation: Backdoors **WITHOUT** label poisoning



ASR with no label poisoning for two poison percentages

Conclusions

- Novel illumination-based video backdoor attack against DNN anti-spoofing detection systems
- The attack is robust against geometric transformation and to some extend against intensity
- With label poisoning, increasing the amplitude and frequency makes the attack more powerful
- Low attack portions are enough

Future Work

- Adapt the backdoor signal to the training set
- Turn the presented attack into a physical attack
- Using physical alteration of the environment

THANK YOU!