SecVID: Correction-based Defense Against Adversarial Video Attacks via Discretization-Enhanced Video Compressive Sensing

Wei Song, Cong Cong, Haonan Zhong, Jingling Xue

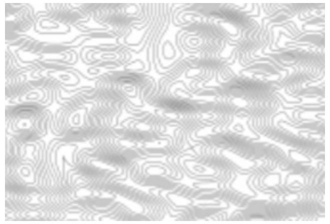
Adversarial Video Examples

The adversary seeks an adversarial x^{adv} of x satisfying:

$$f(x^{adv}) = y_t \qquad \text{if targeted}$$

$$f(x^{adv}) \neq y_0 \qquad \text{if untargeted}$$

x^{adv} is optimized by querying the model until successfully fool the model.



+

ApplyEyeMakeup

→

→ video recognition model → *Playing Def*

Defense Mechanisms

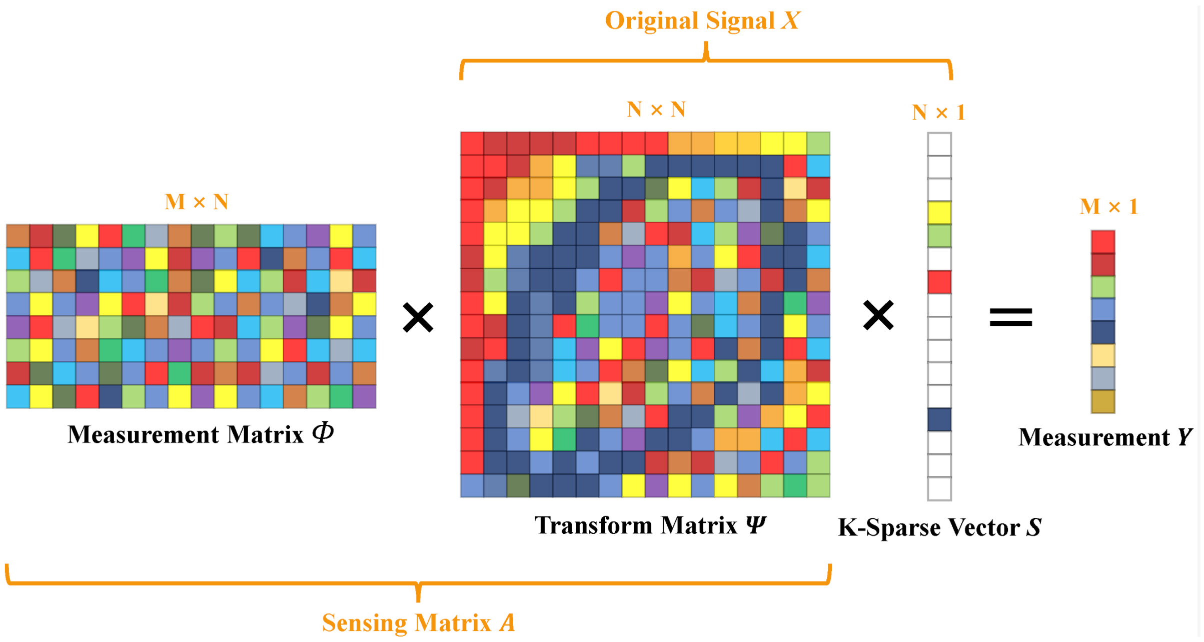
SecVID: an efficient correction-based **video-centric** defense *without accessing video classifiers*, *without requiring known adversarial examples,* and *no need for classifier retraining*

Method	Video-Oriented	Temporal Dynamics Considered	Black-Box	Correction Capability	No Requirement for Prior Adversarial Data	No Model Retraining	No Requirement for Original Dataset
Adversarial Training	✗	✗	✗	✗	✗	✗	✗
Random Smoothing	✗	✗	✗	✗	✓	✗	✗
ComDefend	✗	✗	✓	✓	✓	✓	✗
Compressed&Restore	✗	✗	✗	✓	✗	✓	✗
SESR	✗	✗	✓	✓	✓	✓	✗
DiffPure	✗	✗	✓	✓	✓	✓	✗
FakeDetector	✗	✗	✓	✗	✗	✗	✓
Input Transformations	✗	✗	✓	✗	✓	✓	✗
OUDefend	✓	✓	✗	✗	✓	✓	✗
DP	✓	✓	✗	✗	✗	✓	✗
AdvIT	✓	✓	✓	✗	✓	✓	✗
SecVID	✓	✓	✓	✓	✓	✓	✗

SecVID Key Idea

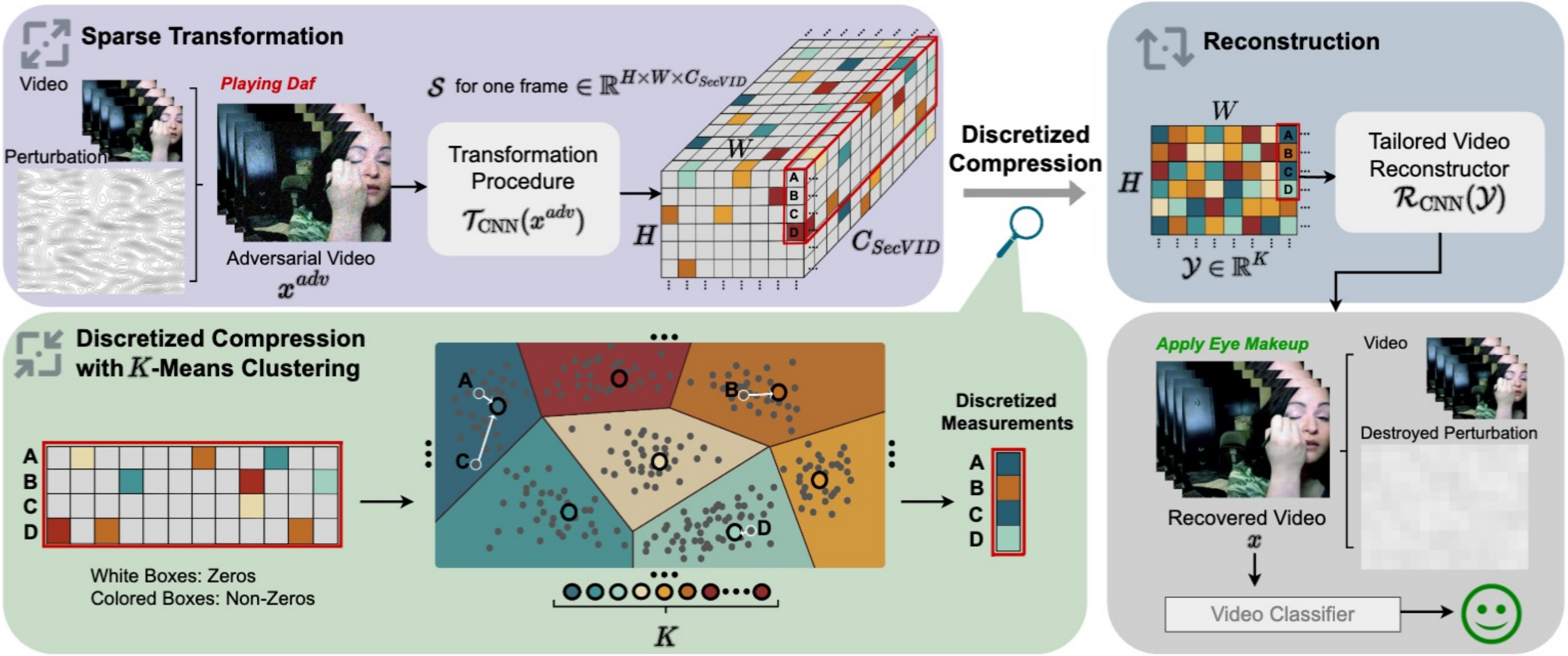
SecVID's key insight lies in its innovative application of compressive sensing, a technique originally for signal compression:

- The signal is **sparse** at some space
- With random **few measurements**, the original signal can be **recovered**



SecVID Overview

- Sparse Transformation
- Discretized Compression
- Reconstruction



SecVID – Sparse Transformation

- ***Sparsity Change (SC).*** *This reflects the variation in the number of significant coefficients resulting from the transformation, where a coefficient is deemed* significant *if its absolute value exceeds a small positive threshold* τ*:*

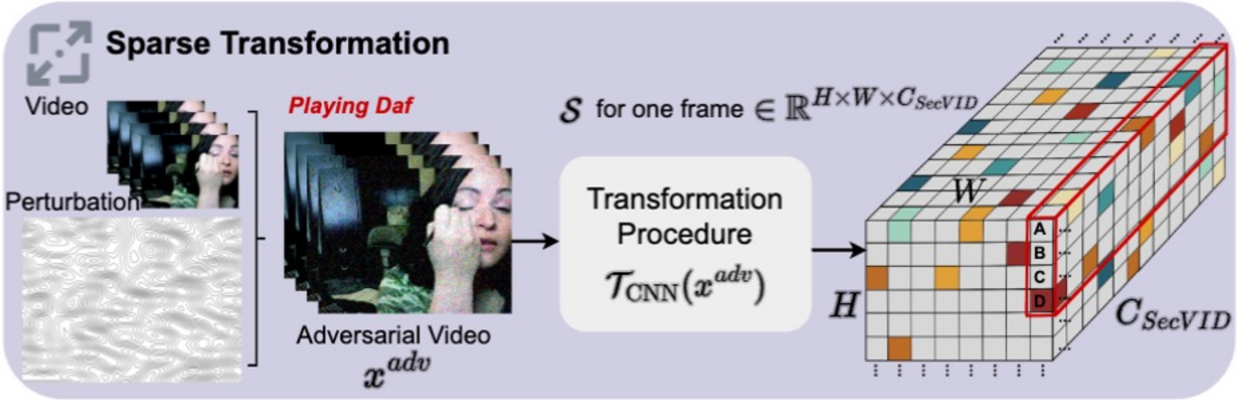
$$SC_{\mathcal{T}}(x^{adv}) = \frac{|\{i \mid |\mathcal{T}(x_i^{adv})| > \tau\}|}{N} - \frac{|\{i \mid |x_i^{adv}| > \tau\}|}{N}$$

- ***Intensity Redistribution (IR).*** *This quantifies the shift in energy or intensity distribution of the transformed signal compared to the original signal:*

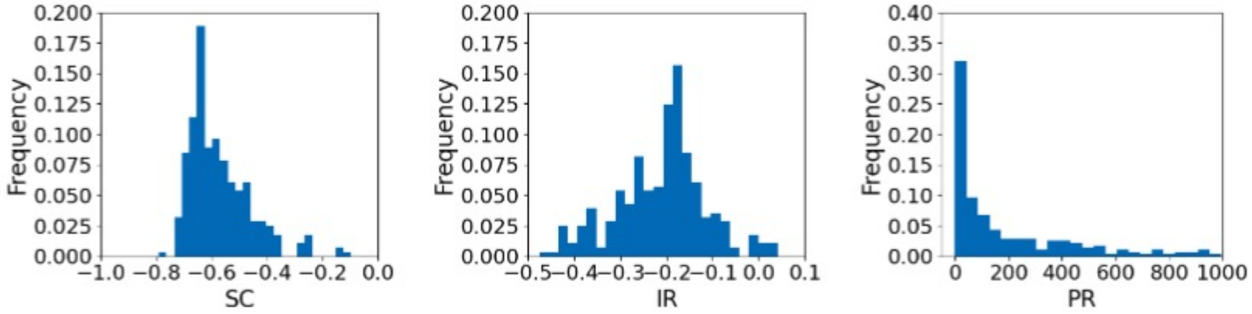
$$IR_{\mathcal{T}}(x^{adv}) = \frac{1}{N}\sum_{i=1}^{N}|\mathcal{T}(x_i^{adv})|^2 - \frac{1}{N}\sum_{i=1}^{N}|x_i^{adv}|^2$$

- ***Positional Redistribution (PR).*** *This metric evaluates the positional shifts of non-zero elements in a signal post-transformation, using, for example, the Wasserstein distance W [32] to calculate the minimal "work" needed to transform one distribution into another. It specifically applies to sets* $P_{x^{adv}}$ *and* $P_{\mathcal{T}(x^{adv})}$*, which represent non-zero elements in* x^{adv} *and* $\mathcal{T}(x^{adv})$*, respectively:*

$$PR_{\mathcal{T}}(x^{adv}) = W(P_{x^{adv}}, P_{\mathcal{T}(x^{adv})})$$



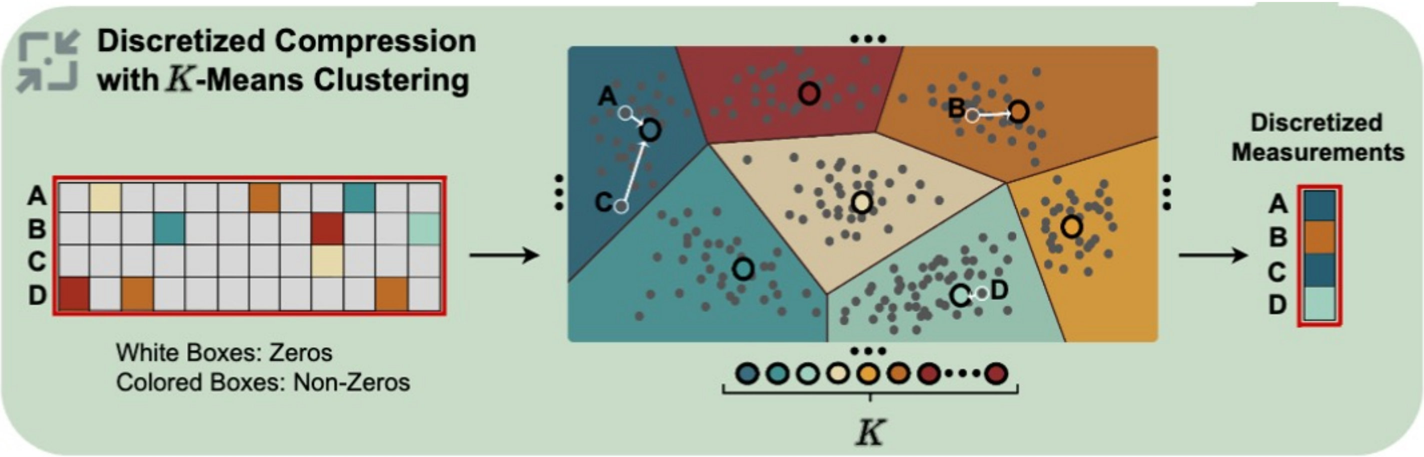
$$DoS = C_{\text{SECVID}}/C$$

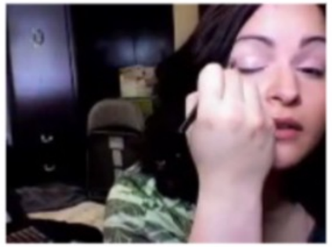


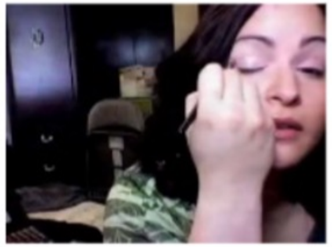
SecVID – Discretized Compression

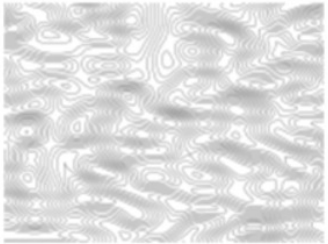
We exploit the inherent continuity of adversarial perturbations—often their Achilles' heel—by employing discretized compression. This process involves discretization, which transforms the data from a smooth **continuum into a distinct, jagged discrete space, and compression**, which further compacts the data, effectively neutralizing perturbations.

We employ the K-Means clustering algorithm for our discretized compression component, chosen for its efficiency in discretizing continuous data and its lightweight characteristics. This method quantizes the continuous sparse representation by assigning each data point to the nearest cluster centroid.




Clean frame


Perturbed frame


Perturbation

Destroyed perturbation

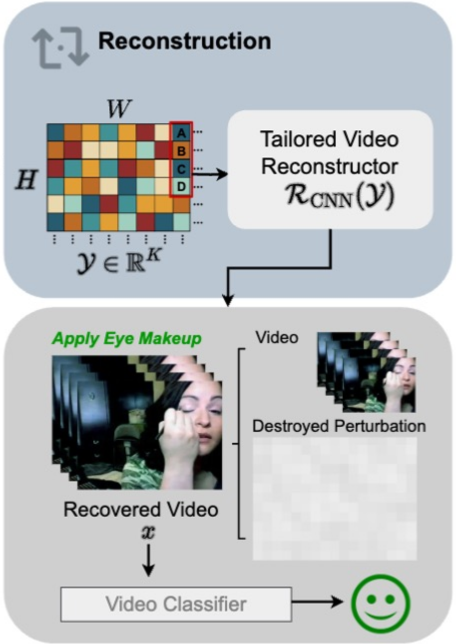
SecVID – Reconstruction

Finally, given the discretized measurements, we can reconstruct the videos. Although SecVID may not completely restore adversarial videos to their original state, it significantly recovers their quality.

It is co-trained with the sparse transformation module, with the total loss defined as:

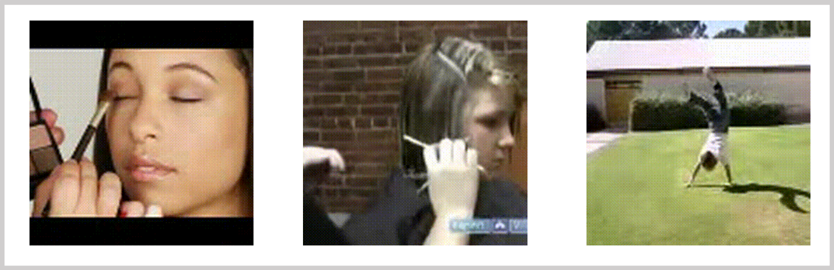
$$L_{\text{loss}} = \alpha L_{\text{cont}} + \beta L_{\text{temp}} + \gamma L_{\text{per}} + \delta L_S$$

- Content loss: $L_{\text{cont}}(x_t, x_t') = \sum_{t=1}^{T} \sum_{l} \frac{1}{H_l W_l C_l} \left\| \phi_l^{\text{VGG-19}}(x_t) - \phi_l^{\text{VGG-19}}(x_t') \right\|_2^2$
- Temporal loss: $L_{\text{temp}}(x_t', x_{t+1}') = \sum_{i=1}^{T-1} \frac{1}{HWC} \left\| x_{t+1}' - \text{warp}(x_t', \Omega^{\text{SpyNet}}(x_t', x_{t+1}')) \right\|_2^2$
- Perceptual loss: $L_{\text{per}}(x_t, x_t') = \sum_{t=1}^{T} \frac{1}{HWC} \left\| x_t - x_t' \right\|_2^2$
- Sparse transformation loss: $L_S = \sum_{t=1}^{T} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{C_{\text{SecVID}}} \| S_{tijk} \|_1$

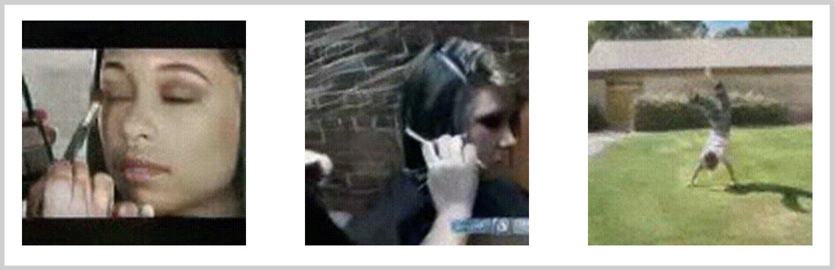


SecVID – Reconstruction

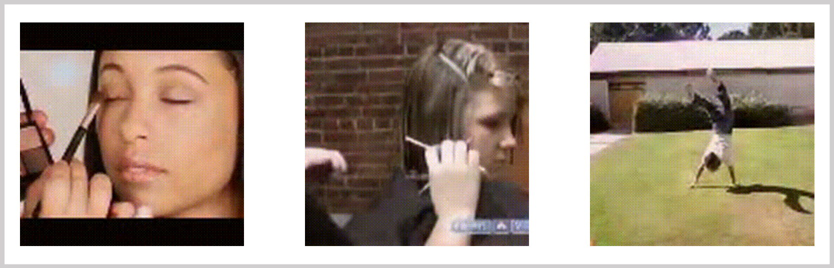
Original Videos



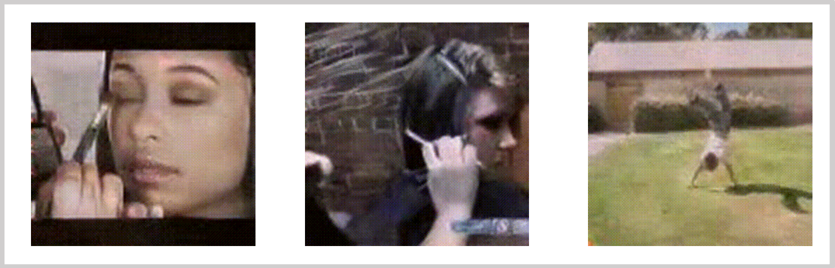
Adversarial Videos



Reconstructed Original Videos



Reconstructed Adversarial Videos



Experiment – Set up

2 Datasets:

- UCF-101
- HMDB-51

2 Video Recognition Models:

- C3D
- I3D

5 Attack Types:

- StyleFool (U), and StyleFool (T) from **StyleFool** *(IEEE S&P'2023)*
- Geo-Trap (U), and Geo-Trap (T) from **Geo-Trap** *(NeurIPS' 2021)*
- U3D (U) from **U3D** *(IEEE S&P'2023)*

1 Adaptive Attack:

- Adversarial Patch Attack

7 baselines:

- Video-centric: AdvIT, OUDefend
- Image-focused: Adversarial Training (AT), Input Transformations (IT), Random Smoothing (RS), ComDefend, DiffPure

Evaluation – Defense Performance

Comparing SecVID with AdvIT, AT, IT, RS, OUDefend, ComDefend, and DiffPure for their **DSRs (Detection Success Rates)** on adversarial videos.

Model	Attack Method	AdvIT		AT		IT		RS		OUDefend		ComDefend		DiffPure		SecVID	
		UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51
C3D	StyleFool (U)	31.2%	25.0%	62.4%	53.1%	8.3%	25.0%	13.8%	6.3%	37.6%	25.0%	42.2%	18.8%	65.1%	59.4%	**83.5%**	**84.3%**
	U3D (U)	12.0%	15.9%	47.3%	38.2%	50.0%	12.5%	16.1%	30.2%	57.1%	32.4%	70.9%	46.0%	71.3%	66.4%	**92.3%**	**83.3%**
	Geo-Trap (U)	19.7%	10.5%	64.5%	58.5%	25.0%	0.0%	23.2%	32.1%	48.7%	22.6%	65.8%	56.6%	63.2%	66.0%	**100.0%**	**100.0%**
	StyleFool (T)	17.5%	15.4%	53.2%	46.9%	20.4%	11.2%	11.3%	8.4%	43.6%	29.4%	54.1%	45.4%	68.6%	53.8%	**79.2%**	**81.8%**
	Geo-Trap (T)	16.7%	19.6%	61.9%	60.8%	0.0%	0.0%	14.3%	25.5%	48.8%	19.6%	63.1%	51.0%	58.3%	64.7%	**83.3%**	**100.0%**
I3D	StyleFool (U)	13.7%	18.6%	57.0%	59.2%	16.7%	11.1%	9.0%	7.3%	43.1%	18.4%	75.0%	63.1%	60.4%	55.8%	**88.9%**	**83.0%**
	U3D (U)	12.3%	9.6%	49.8%	41.2%	20.2%	26.2%	25.2%	29.2%	34.9%	28.6%	18.3%	22.6%	68.1%	67.4%	**83.1%**	**94.4%**
	Geo-Trap (U)	19.7%	9.2%	64.9%	61.9%	38.1%	0.0%	21.6%	27.0%	28.9%	23.8%	82.5%	63.5%	53.6%	57.1%	**91.8%**	**88.8%**
	StyleFool (T)	15.5%	13.9%	57.3%	53.8%	2.8%	3.0%	10.5%	6.3%	39.2%	13.6%	38.5%	43.9%	58.7%	65.7%	**96.5%**	**83.3%**
	Geo-Trap (T)	11.9%	9.8%	58.1%	61.0%	41.9%	0.0%	18.9%	23.7%	24.3%	20.3%	85.1%	33.9%	62.2%	67.8%	**83.8%**	**83.1%**

Evaluation – Managing Clean Videos

Comparing SecVID with AdvIT, AT, IT, RS, OUDefend, ComDefend, and DiffPure in **managing clean videos** from UCF-101 and HMDB-51, using false positive rate for AdvIT (detection-only), and accuracy for the others (protection-oriented).

Model	False Positive Rates AdvIT		Recognition Accuracy Unprotected		AT		IT		RS		OUDefend		ComDefend		DiffPure		SECVID	
	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51
C3D	3.2%	4.1%	78.3%	60.2%	76.2%	59.2%	49.7%	38.2%	65.6%	51.2%	59.7%	48.4%	69.3%	55.7%	71.5%	52.3%	73.7%	56.9%
I3D	3.2%	4.1%	87.6%	62.5%	87.3%	62.5%	58.4%	42.3%	53.8%	54.3%	55.3%	42.9%	78.6%	56.3%	82.9%	56.4%	85.2%	60.3%

Evaluation – Security Costs

Average inference time (ms) per video of classifiers protected by SecVID, AdvIT, OUDefend, ComDefend, DiffPure, and “Unprotected” (representing AT, IT, and RS), with 300 clean videos randomly selected from each of UCF-101 and HMDB-51.

Model	Unprotected		AdvIT		OUDefend		ComDefend		DiffPure		SecVID	
	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51
C3D	6.21	6.99	434.17**(69.91×)**	436.29**(62.42×)**	11.93**(1.92×)**	18.76**(2.68×)**	12.50**(2.01×)**	23.69**(3.39×)**	458.40**(73.82×)**	446.24**(63.86×)**	13.72**(2.20×)**	30.60**(4.40×)**
I3D	10.75	8.40	438.71**(40.81×)**	437.70**(52.11×)**	19.58**(1.82×)**	24.63**(2.93×)**	17.04**(1.59×)**	25.10**(2.99×)**	462.94**(43.06×)**	447.65**(53.29×)**	18.34**(1.71×)**	34.12**(4.06×)**

Evaluation – Adaptive Attack

We assess SecVID's robustness against adaptive attacks targeting its sparse transformation using **adversarial patch attack**. Adversarial patch attack exploits spatial sparsity by perturbing a strategically selected small area in each image/frame.

We demonstrate that although such sparse attacks are problematic due to their **human-perceptibility**, SecVID purposely designed to counter human-imperceptible perturbations, effectively mitigates these attacks through its discretized compression strategy.

Examples of adversarial patch attack from "Adversarial Patch"

Defense	C3D		I3D	
	UCF-101	HMDB-51	UCF101	HMDB-51
SecVID	82.7%	74.7%	89.3%	72.0%
DiffPure	69.3%	72.0%	66.7%	65.3%

Evaluation – Ablation Study

Impact of ***DoS*** on SecVID's DSR, evaluated with four *DoS* levels

Model	Attack Method	$DoS = 1$		$DoS = 2$		$DoS = 3$		$DoS = 4$	
		UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51
C3D	StyleFool (U)	69.7%	75.0%	72.5%	71.9%	74.3%	75.0%	**83.5%**	**84.3%**
	U3D (U)	74.9%	50.0%	84.6%	66.7%	88.5%	68.8%	**92.3%**	**83.3%**
	Geo-Trap (U)	75.0%	**100.0%**	87.5%	**100.0%**	82.9%	**100.0%**	**100.0%**	**100.0%**
	StyleFool (T)	75.2%	73.4%	73.8%	74.8%	76.6%	78.3%	**79.2%**	**81.8%**
	Geo-Trap (T)	70.0%	**100.0%**	71.1%	**100.0%**	77.8%	**100.0%**	**83.3%**	**100.0%**
I3D	StyleFool (U)	87.5%	68.9%	86.1%	**83.5%**	**89.0%**	**83.5%**	88.9%	83.0%
	U3D (U)	69.8%	86.4%	81.7%	83.1%	78.1%	**89.4%**	83.1%	**94.4%**
	Geo-Trap (U)	87.6%	66.7%	89.7%	66.7%	89.7%	66.7%	**91.8%**	**88.8%**
	StyleFool (T)	**100.0%**	78.8%	88.1%	78.3%	90.9%	80.3%	96.5%	**83.3%**
	Geo-Trap (T)	71.6%	66.1%	81.1%	66.1%	**83.8%**	66.1%	**83.8%**	**83.1%**

Evaluation – Ablation Study

Impact of varying **cluster counts** on SecVID in terms of DSR, evaluated across four distinct cluster counts

Model	Attack Method	Without Discretized Compression		With Discretized Compression							
				$K = 128$		$K = 256$		$K = 512$		$K = 1024$	
		UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51
C3D	StyleFool (U)	63.3%	40.0%	79.8%	78.1%	82.5%	75.0%	80.0%	81.3%	**83.5%**	**84.3%**
	U3D (U)	50.0%	54.8%	83.3%	85.5%	83.3%	84.7%	91.6%	**86.7%**	**92.3%**	83.3%
	Geo-Trap (U)	87.5%	95.5%	75.0%	**100.0%**	75.0%	**100.0%**	87.5%	**100.0%**	**100.0%**	**100.0%**
	StyleFool (T)	64.8%	61.2%	**79.2%**	73.3%	**79.2%**	75.2%	**79.2%**	76.1%	**79.2%**	**81.8%**
	Geo-Trap (T)	78.0%	66.7%	73.7%	**100.0%**	73.7%	**100.0%**	**83.8%**	66.7%	83.3%	**100.0%**
I3D	StyleFool (U)	85.4%	63.1%	86.1%	61.5%	**90.3%**	66.7%	88.9%	67.4%	88.9%	**83.0%**
	U3D (U)	59.8%	89.5%	69.4%	72.2%	**83.3%**	80.8%	**83.3%**	83.3%	83.1%	**94.4%**
	Geo-Trap (U)	83.8%	66.7%	**93.5%**	66.7%	**93.5%**	74.2%	90.3%	78.5%	91.8%	**88.8%**
	StyleFool (T)	84.7%	73.5%	94.4%	79.5%	94.4%	79.5%	94.4%	81.8%	**96.5%**	**83.3%**
	Geo-Trap (T)	78.4%	66.1%	73.0%	62.1%	73.0%	65.8%	**83.8%**	66.1%	**83.8%**	**83.1%**

Evaluation – Ablation Study

Impact of **sparse transformation loss** on SecVID's DSR

Model	Attack Method	Without L_S		With L_S	
		UCF-101	HMDB-51	UCF-101	HMDB-51
C3D	StyleFool (U)	74.3%	81.3%	83.5%	84.3%
	U3D (U)	84.9%	81.1%	92.3%	83.3%
	Geo-Trap (U)	81.6%	86.8%	100.0%	100.0%
	StyleFool (T)	75.0%	78.3%	79.2%	81.8%
	Geo-Trap (T)	78.6%	86.3%	83.3%	100.0%
I3D	StyleFool (U)	75.0%	68.4%	88.9%	100.0%
	U3D (U)	81.4%	84.4%	83.1%	94.4%
	Geo-Trap (U)	86.7%	81.0%	91.8%	88.8%
	StyleFool (T)	93.7%	79.8%	96.5%	83.3%
	Geo-Trap (T)	78.4%	76.3%	83.8%	83.1%

Evaluation – More in Paper

- Reconstruction quality evaluation (SSIM, PSNR, and FID)
- SecVID's performance under various perturbation intensities
- SecVID's security costs with various settings

Conclusion

- A novel correction-based adversarial **video** defense framework build on video compressive sensing theory
- A **discretized compression** technique to mitigate adversarial perturbations
- Enhancing video recognition security introduces **trade-offs**, such as slightly reduced recognition accuracy and longer inference times.
- These trade-offs pinpoint future research directions, especially in **cost-effective** adversarial video defense methods that selectively/negligibly impact security

Thanks
Q&A

Contact wei.song1@unsw.edu.au for any further questions