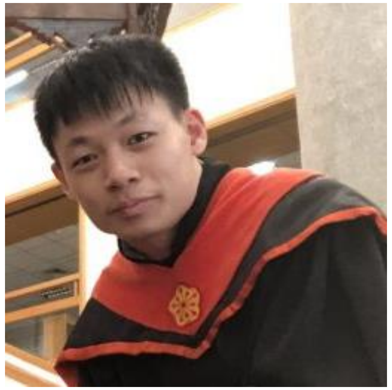
MultAV: Multiplicative Adversarial Videos

AVSS 2021



Shao-Yuan Lo and Vishal M. Patel

Johns Hopkins University

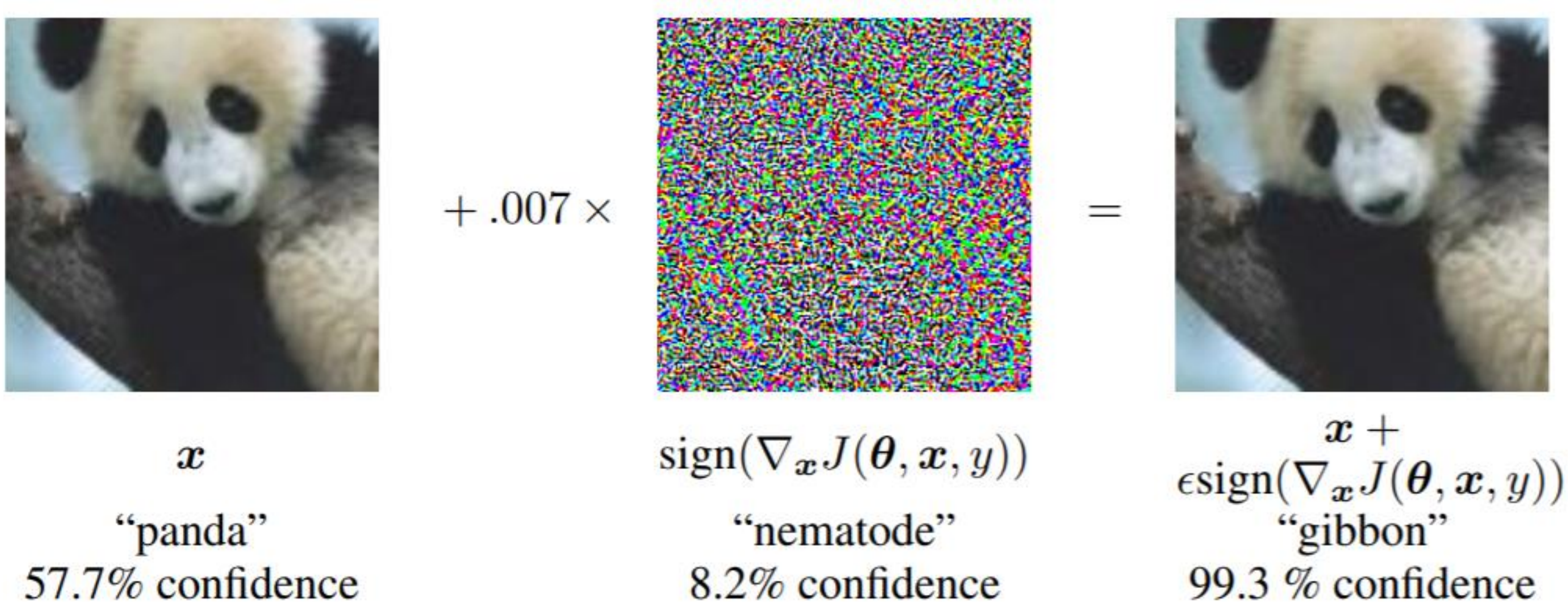
Recall: Adversarial Examples

$$x_{adv} = x + \delta$$

$$f(\boldsymbol{x}_{adv}) \neq y$$

Recall: Adversarial Examples

- Deep networks are **vulnerable** to adversarial examples.



Goodfellow et al. Explaining and Harnessing Adversarial Examples. ICLR'15.

Adversarial Videos

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

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

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

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

Multiplicative Adversarial Videos

- **Additive** attack:

$$x_{adv} = x + \delta$$

- **Multiplicative** attack:

$$x_{adv} = x \odot \delta$$

Multiplicative Adversarial Videos

- **Add-L∞:**

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

$$|x^{adv} - x| \leq \epsilon$$

- **Mult-L∞:**

$$x^{adv} = x \odot \alpha^{sign(\nabla_x L(x,y;\theta))}$$

$$\max\left(\frac{x^{adv}}{x}, \frac{x}{x^{adv}}\right) \leq \epsilon$$

Ratio bound

Multiplicative Adversarial Videos

- **Add-L2:**

$$x^{adv} = x + \alpha \cdot \frac{\nabla_x L(x,y;\theta)}{\|\nabla_x L(x,y;\theta)\|_2}$$

$$\| x^{adv} - x \|_2 \leq \epsilon$$

- **Mult-L2:**

$$x^{adv} = x \odot \alpha^{\frac{\nabla_x L(x,y;\theta)}{\|\nabla_x L(x,y;\theta)\|_2}}$$

$$\| \frac{x^{adv}}{x} \|_2 \leq \epsilon$$

Ratio bound

Signal-dependent Perturbation

- **Mult-L∞:**

$$x^{adv} = x \odot \alpha^{sign(\nabla_x L(x,y;\theta))}$$

$$\Updownarrow$$

$$x^{adv} = x + [x \odot \left(\alpha^{sign\left(\nabla_x L(x,y;\theta)\right)} - 1\right)]$$

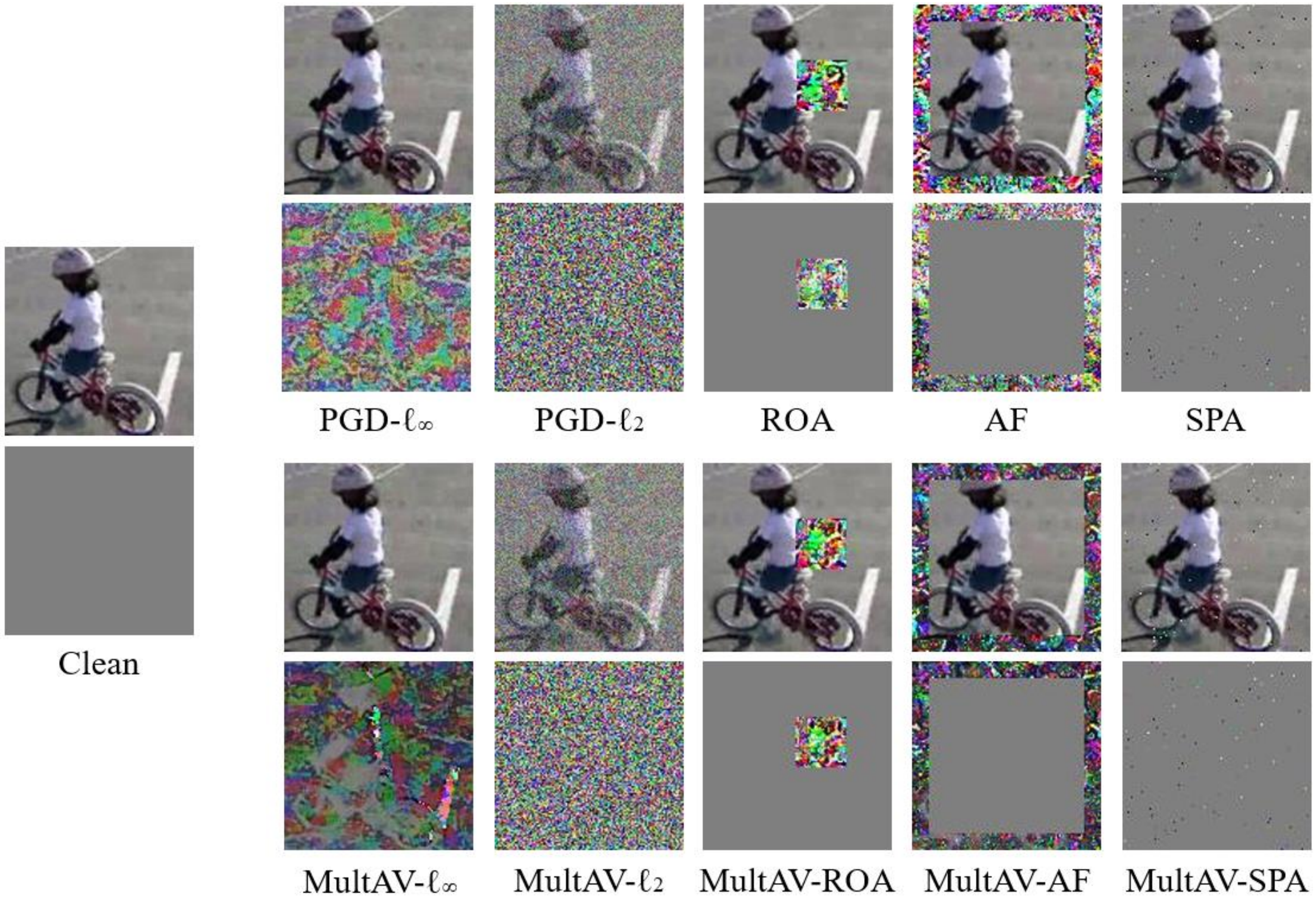
- **Mult-L2:**

$$x^{adv} = x \odot \alpha^{\frac{\nabla_x L(x,y;\theta)}{\|\nabla_x L(x,y;\theta)\|_2}}$$

$$\Updownarrow$$

$$x^{adv} = x + [x \odot \left(\alpha^{\frac{\nabla_x L(x,y;\theta)}{\|\nabla_x L(x,y;\theta)\|_2}} - 1\right)]$$

Visual Results

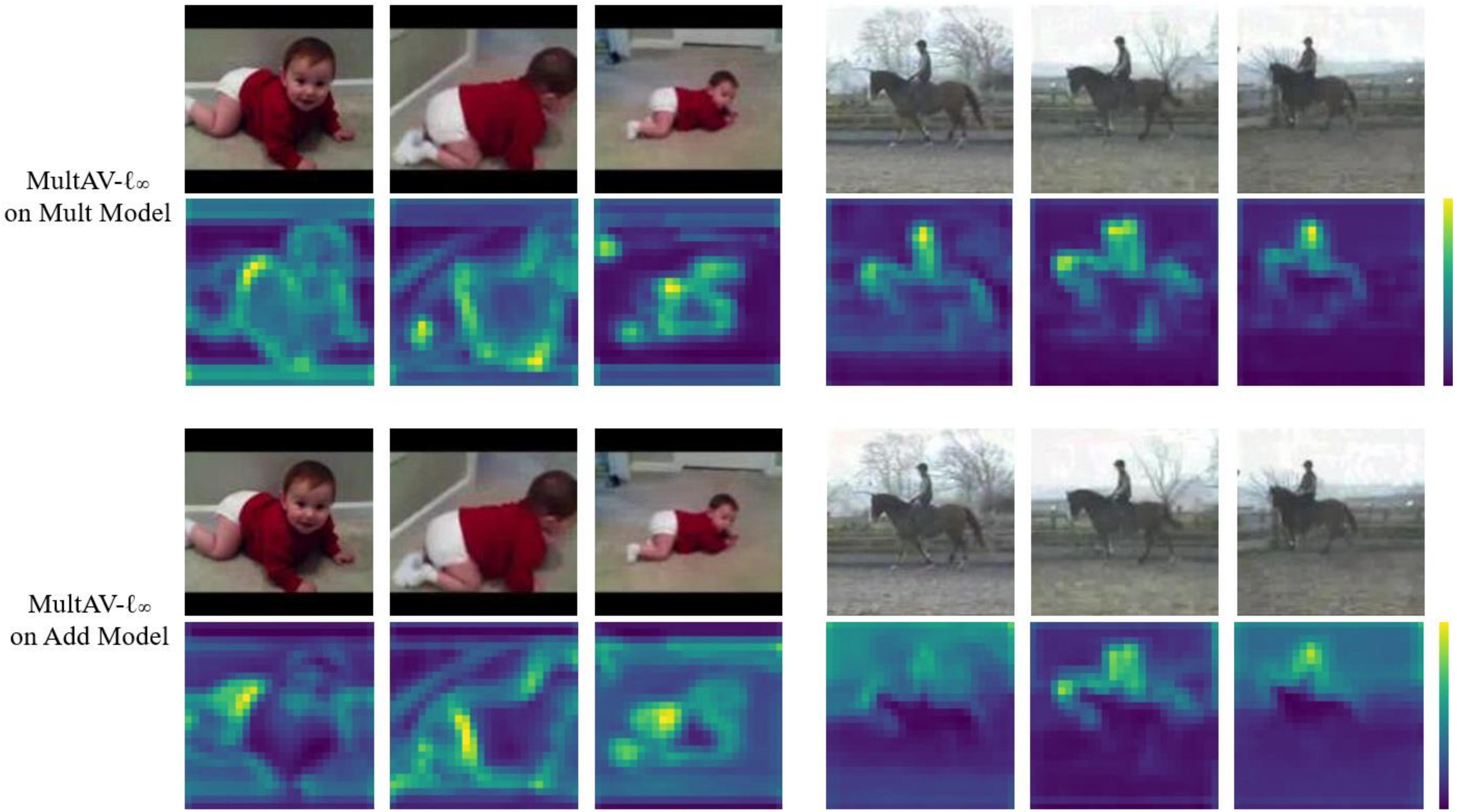


Quantitative Results

- Dataset: UCF-101

Network	Clean	Training	MultAV-ℓ_∞	MultAV-ℓ_2	MultAV-ROA	MultAV-AF	MultAV-SPA
3D ResNet-18	76.90	Clean	7.19	2.67	2.30	0.26	4.02
3D ResNet-18	76.90	Mult	47.00	16.23	44.12	66.35	55.54
		Add	41.61	9.94	42.45	51.23	54.74
			(-5.39)	**(-6.29)**	**(-1.67)**	**(-15.12)**	**(-0.80)**
3D ResNet-18 + 3D Denoise	70.82	Mult	42.69	14.75	39.31	60.53	48.37
		Add	31.46	9.15	37.72	48.98	48.06
			(-11.23)	**(-5.60)**	**(-1.59)**	**(-11.55)**	**(-0.31)**
3D ResNet-18 + 2D Denoise	69.47	Mult	41.87	14.04	40.34	58.97	47.48
		Add	30.16	10.23	39.65	47.82	47.18
			(-11.71)	**(-3.81)**	**(-0.69)**	**(-11.15)**	**(-0.30)**

Feature Visualization



Conclusion

- Propose a new attack method against video recognition networks: Multiplicative Adversarial Videos (MultAV).
- MultAV can generalize to not only Lp-norm attacks, but also different types of physically realizable attacks.
- MultAV challenges the defense approaches that tailored to resisting additive adversarial attacks. We hope to encourage the research community to look into more general and more powerful defense solutions for video recognition networks.