LPS | LABORATORY FOR PHYSICAL SCIENCES

Booz | Allen | Hamilton

ICML 2022

Deploying Convolutional Networks on Untrusted Platforms Using 2D Holographic Reduced Representations

Mohammad Mahmudul Alam, Edward Raff, Tim Oates, James Holt

Objectives

- Deploy on untrusted platform
- Prevent data and model theft
- Fast, heuristic, pseudo-encryption strategy called Connectionist Symbolic Pseudo Secrets (CSPS)
- Neural network with a pseudo-encryption style defense using Neuro-symbolic representation

HRR

$$\mathcal{B} = bind(\text{shape}, \text{square}\blacksquare) + bind(\text{color}, \text{red})$$

$$unbind(\mathcal{B}, \text{shape}) \approx \text{square}$$

$$bind(\mathbf{x}, \mathbf{y}): \mathcal{F}^{-1}(\mathcal{F}(\mathbf{x}) \odot \mathcal{F}(\mathbf{y}))$$

$$unbind: \; y^{\tau} = \mathcal{F}^{-1}\left(\frac{1}{\mathcal{F}(y)}\right)$$

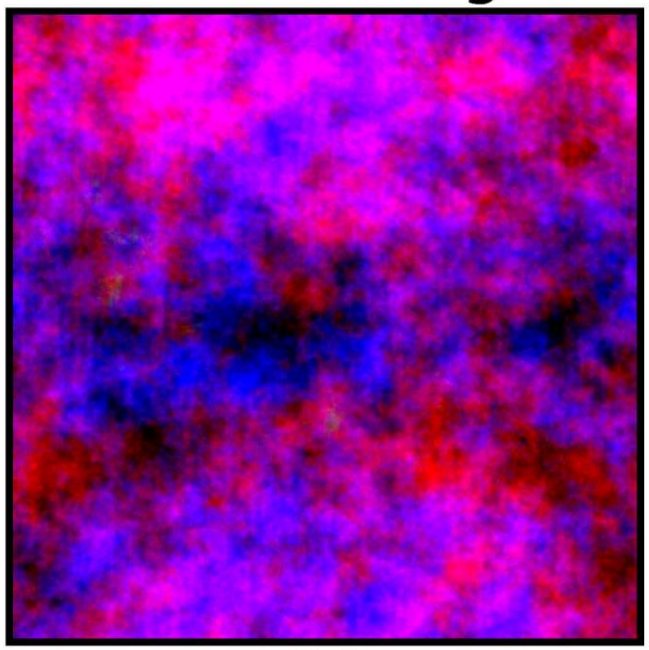
$$projection\ \pi(x): \mathcal{F}^{-1}\left(\frac{\mathcal{F}(x)}{|\mathcal{F}(x)|}\right)$$

2D HRR

Original Image

(a)

Bound Image

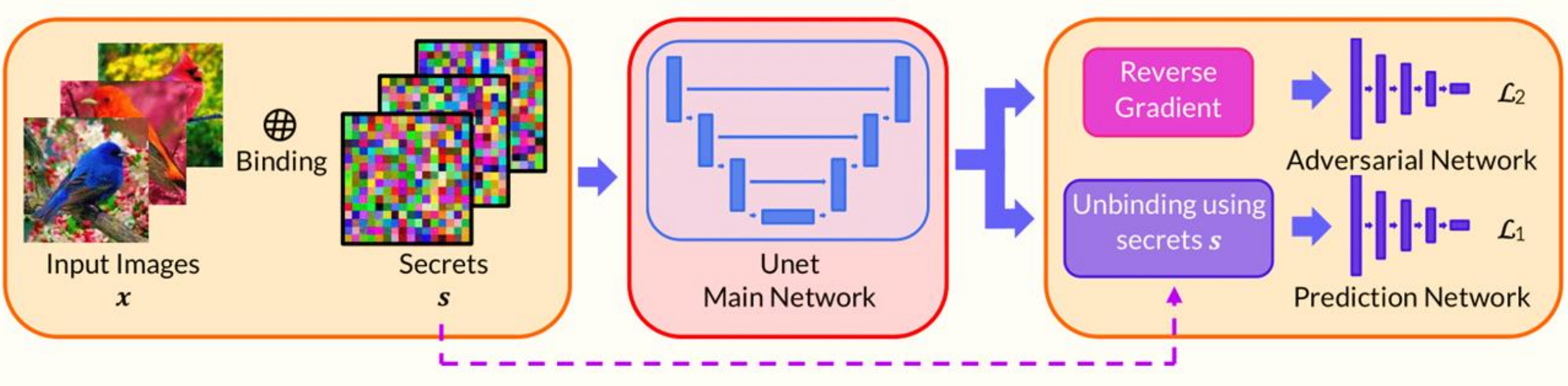


(b)

Retrieved Image

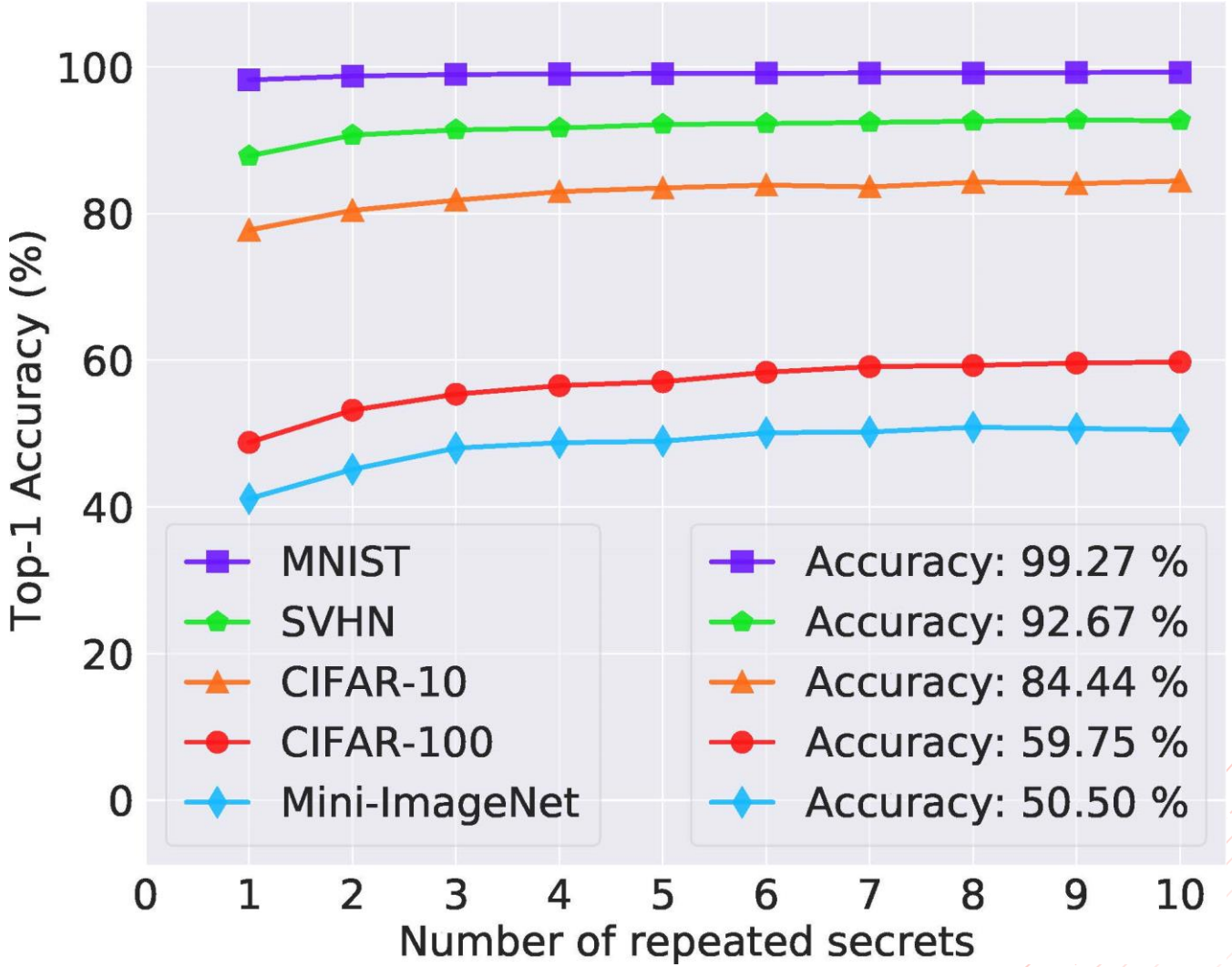
(c)

CSPS Network



Accuracy Results

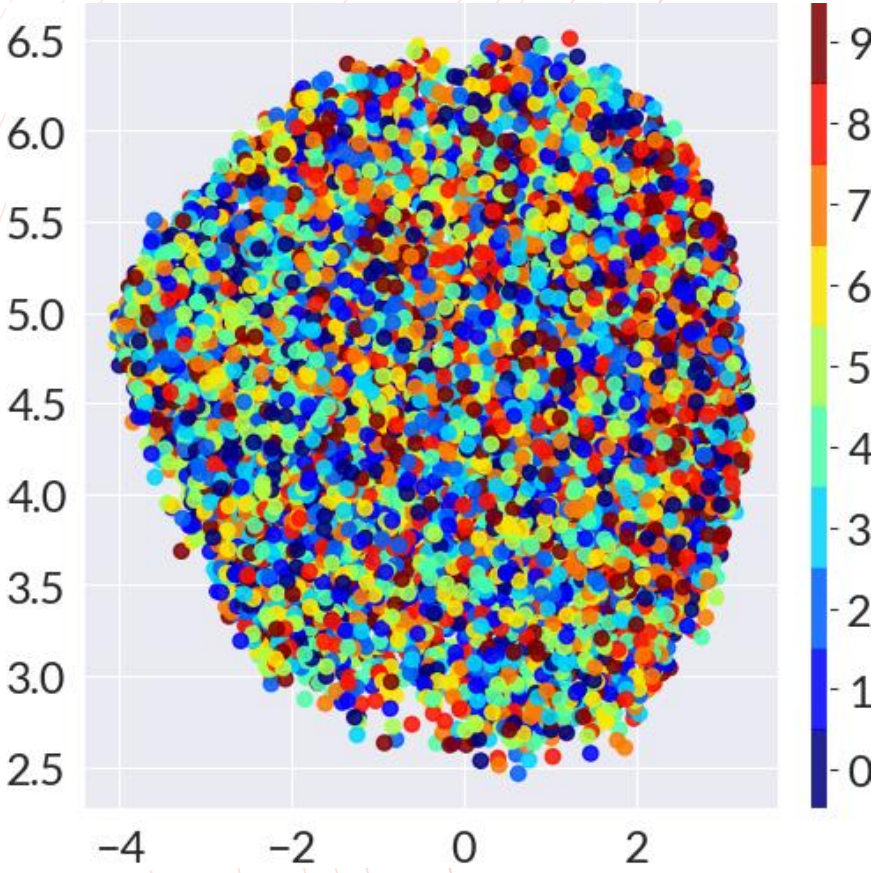
Dataset	Model	Top-1	Top-5
MNIST 28×28	Base	98.80	–
	CSPS	98.51	–
SVHN 32×32	Base	93.76	–
	CSPS	88.44	–
CIFAR-10 32×32	Base	83.57	–
	CSPS	78.21	–
CIFAR-100 32×32	Base	62.59	86.99
	CSPS	48.84	75.82
Mini-ImageNet 84×84	Base	55.73	80.55
	CSPS	40.99	66.99



Run-time Results

Dataset	Our CSPS	HE Est.
MNIST	4.56 Seconds	2 Hours 46 Minutes
SVHN	12.44 Seconds	55 Hours 32 Minutes
CIFAR-10	7.58 Seconds	21 Hours 20 Minutes
CIFAR-100	9.07 Seconds	43 Hours 53 Minutes
Mini-ImageNet	28.37 Seconds	Timeout

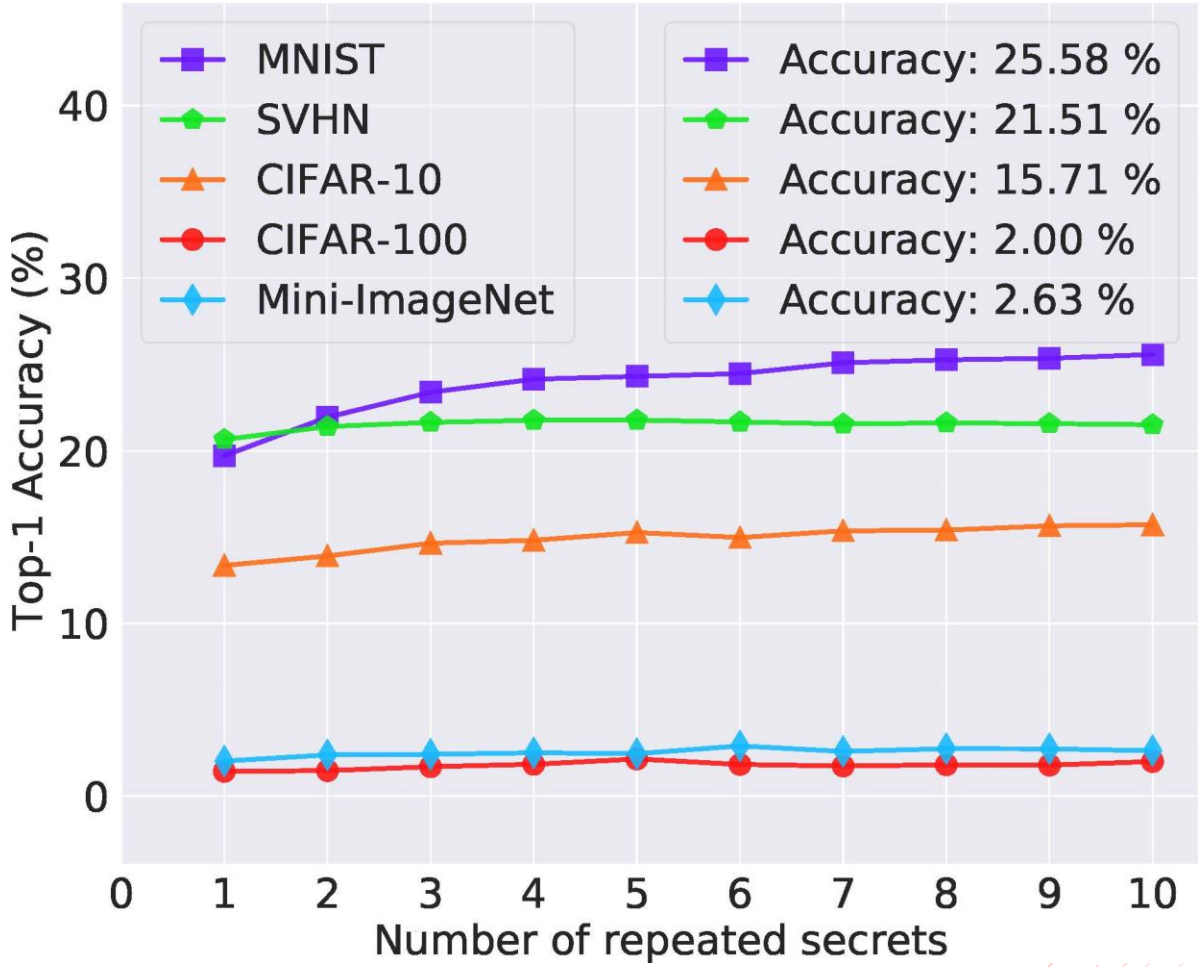
Realistic Adversary



	MNIST	SVHN	CIFAR 10	CIFAR 100	Mini ImgNet
K-Means	1.28	0.06	0.21	0.03	0.08
Spectral	0.01	0.01	0.00	0.00	0.02
GMM	1.28	0.06	0.17	0.04	0.09
Birch	1.51	0.03	0.13	0.05	0.07
HDBSCAN	0.00	0.00	0.00	0.00	0.00
K-Means	-0.02	-0.01	0.18	0.54	0.42
GMM	0.01	0.00	0.09	0.61	0.44
Birch	0.20	0.00	0.14	0.45	0.35
HDBSCAN	0.00	-0.24	1.23	0.01	0.02

Overly Strong Adversary

Dataset	Top-1 Accuracy (%)	Top-5 Accuracy (%)
MNIST	19.72	–
SVHN	21.13	–
CIFAR-10	12.91	–
CIFAR-100	2.66	10.33
Mini-ImageNet	4.68	15.01

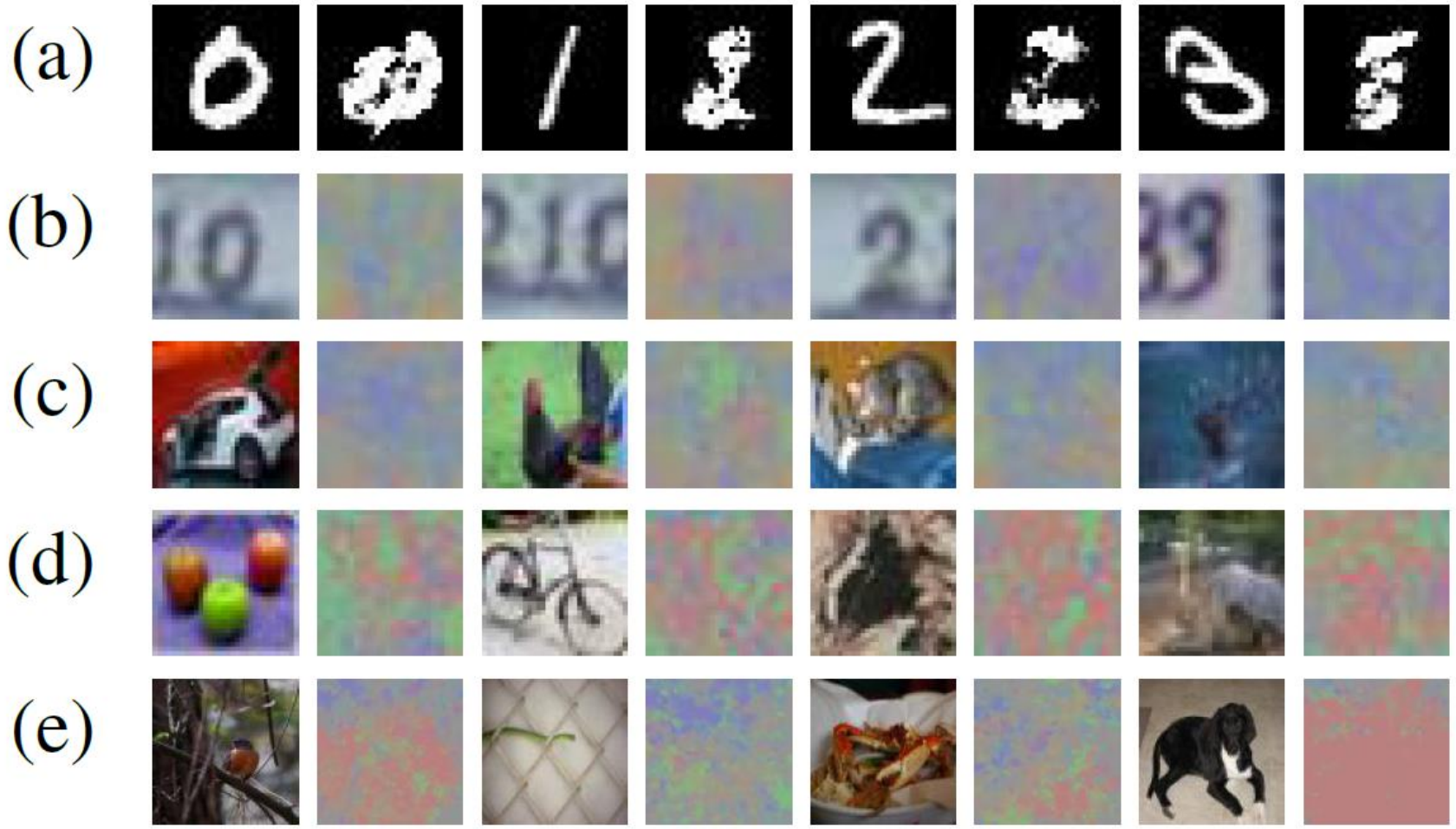


Model Inversion Adversary

Model inversion attack by the adversary using projected gradient descent to unbind the bound images given sample of the original images

Model Inversion Adversary Contd.

Inversion attack using a trained auto-encoder to directly unbind bound images



Conclusive Remarks

- Majority of the computation is done in the untrusted platform
- Compared to SOTA our CSPS is $\approx 5000\times$ faster and sends $\approx 18{,}000\times$ less data per query
- Not provably secure, should be used with caution
- Scalability is limited

Thanks

Any Questions?

Poster at **Hall E** from 6 pm to 8 pm