Defense Against Adversarial Attacks

Recall the classical detection methods

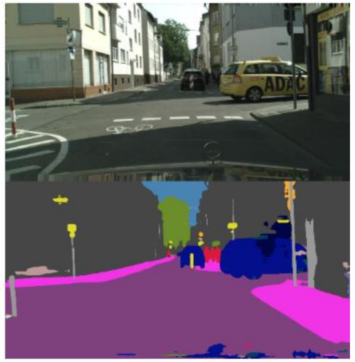
- Pre-processing the image with different transformation methods
- Train a network to tell the adversarial instances apart
- Leverage spatial/temporal properties to check the consistency – indicating adversarial behaviors
- Map the data to other data manifold by computing meaningful metric to measure differences

Beyond the Min-max Game

- What if we have more knowledge about our learning tasks?
  - Properties of learning tasks and data
  - General understanding about ML models

Characterize Adversarial Examples Based on Spatial Consistency Information for Semantic Segmentation

- Attacks against semantic segmentation
  - State-of-the-art attacks against segmentation: Houdini [NIPS2017], DAG [ICCV 2017]
  - We design diverse adversarial targets: hello kitty, pure color, a real scene, ECCV, color shift, strips of even color of classes
  - Cityscapes and BDD datasets



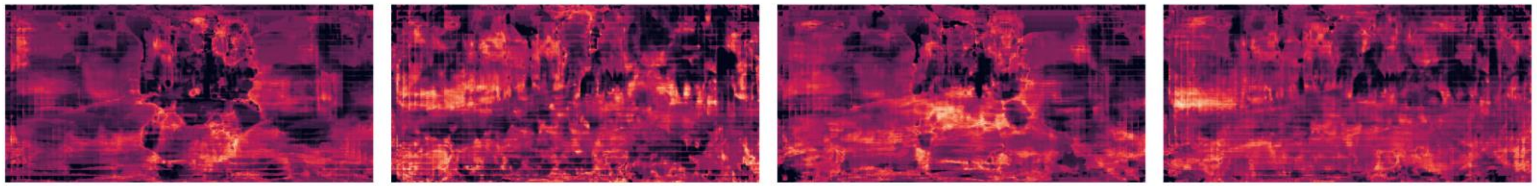
Benign

Adversarial Examples

Spatial Context Information

- Spatial consistency is a distinct property of image segmentation
- Perturbation at one pixel will potentially affect the prediction of surrounding pixels $\mathcal{H}(m) = -\sum_j \mathcal{V}_m[j] \log \mathcal{V}_m[j]$

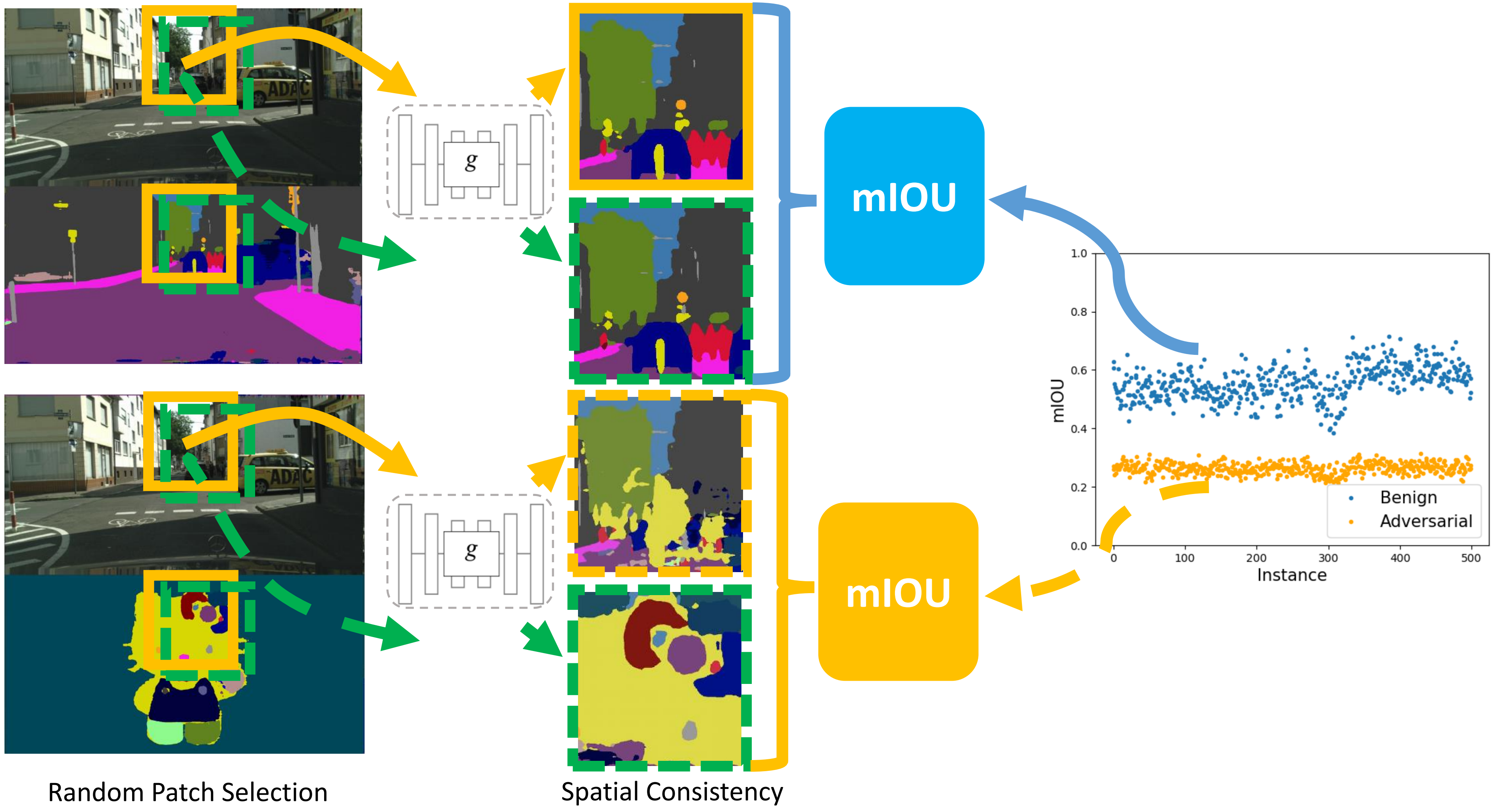
(a) Benign example (b) Heatmap of benign image



(c) DAG | Kitty (d) DAG | Pure (e) Houdini | Kitty (f) Houdini | Pure

Perturbation on single patch may loss its adversarial effect

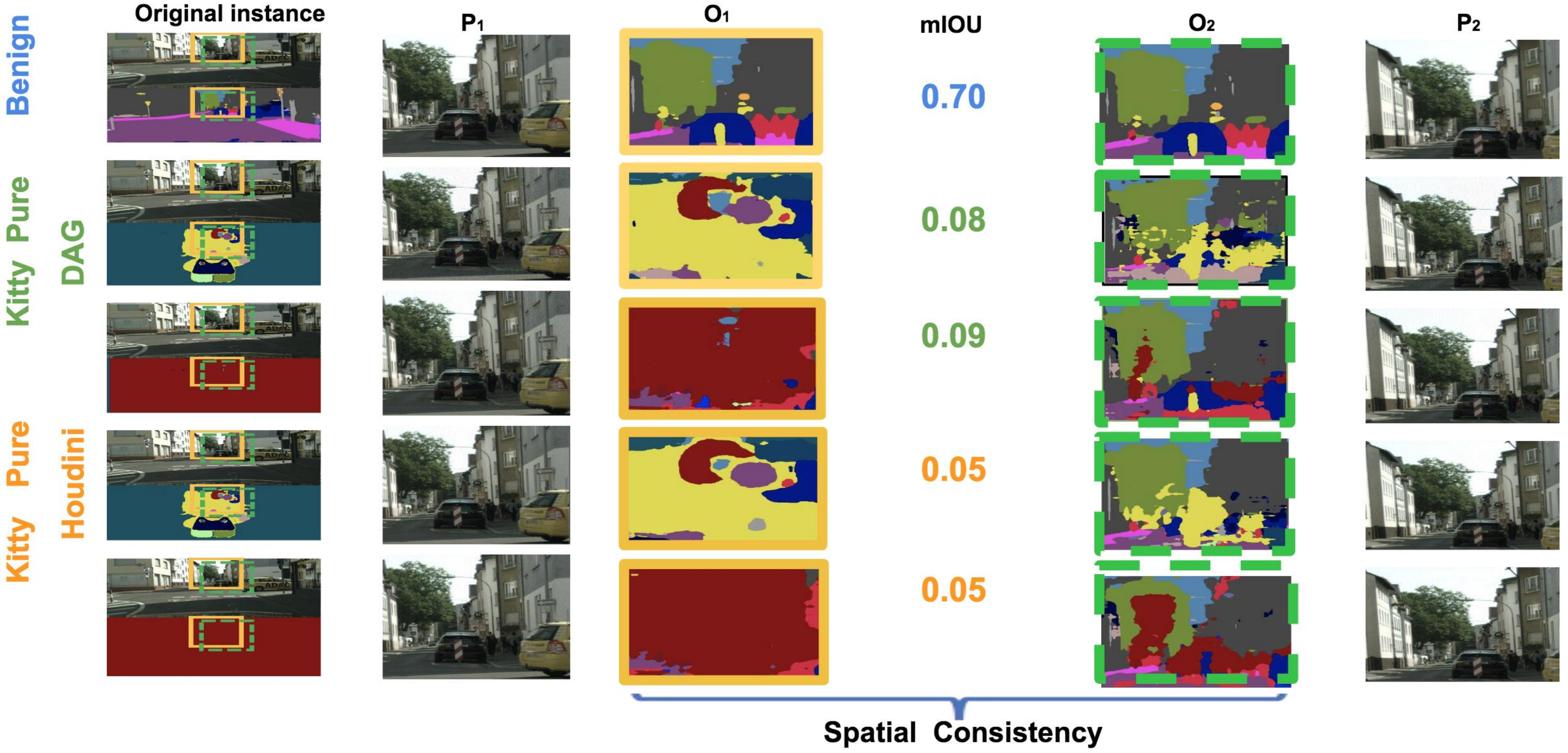
- Spatial consistency: the consistency of segmentation results for randomly selected patches from an image
- Such spatial consistency information from benign and adversarial instances are distinguishable
- We apply mIOU to compare the segmentation results between patches
  - For each class, Intersection over Union (IOU) is calculated as TP/(TP+FP+FN). Here we calculate the relative mIOU for each pair of patches



Pipeline of spatial consistency based detection for adversarial examples on semantic segmentation

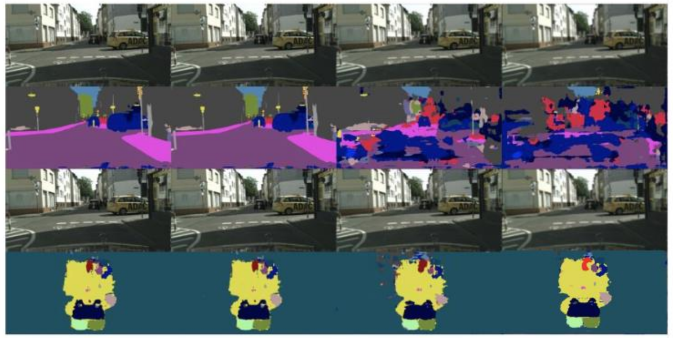
We apply mIOU to evaluate the consistency information for patches from benign and adversarial instances quantitatively

- Detection

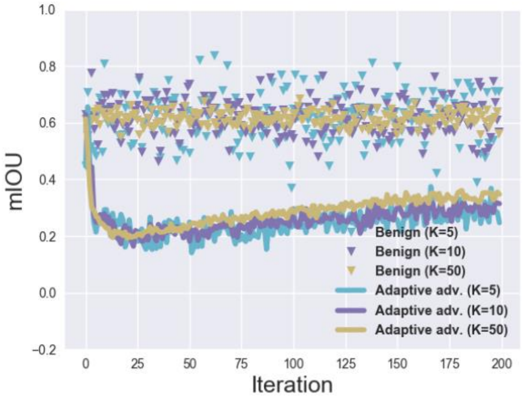


Adaptive Attack Against Spatial Consistency Based Detection

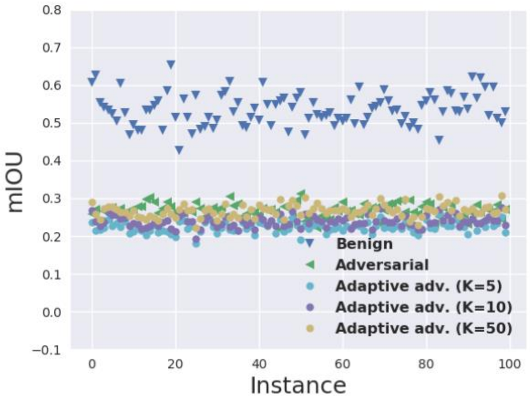
- Adaptive attack:
  - Assume the attacker has perfect knowledge of #selected patches: K
  - We generate perturbation that the selected k patches can all be mis-segmented to the corresponding regions within adversarial target



(a) Adaptive attack against image scaling



(b) Convergence analysis of adaptive attack



(c) Adaptive attack against spatial consistency

Detecting adversarial instances based on spatial consistency information

- Both the spatial consistency based detection and the scaling based baseline achieve promising detection rate on different attacks
- The scaling based baseline fails to detect strong adaptive attacks while the spatial based method can

Method		Model	mIOU	Detection DAG Pure	Detection DAG Kitty	Detection Houdini Pure	Detection Houdini Kitty	Detection Adap DAG Pure	Detection Adap DAG Kitty	Detection Adap Houdini Pure	Detection Adap Houdini Kitty
Scale (std)	0.5	DRN (16.4M)	66.7	100%	95%	100%	99%	100%	67%	100%	78%
	3.0			100%	100%	100%	100%	100%	0%	97%	0%
	5.0			100%	100%	100%	100%	100%	0%	71%	0%
Spatial (K)	1	DRN (16.4M)	66.7	91%	91%	94%	92%	98%	94%	92%	94%
	5			100%	100%	100%	100%	100%	100%	100%	100%
	10			100%	100%	100%	100%	100%	100%	100%	100%
	50			100%	100%	100%	100%	100%	100%	100%	100%

Takeaways

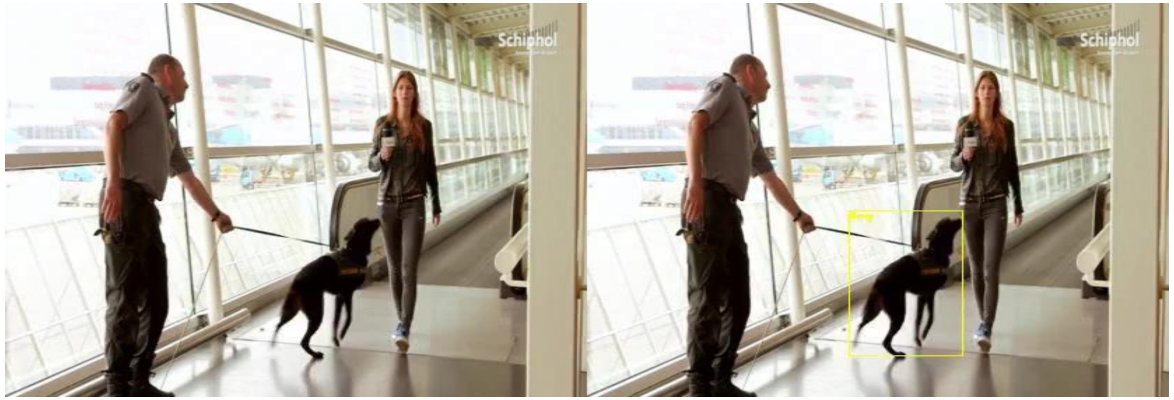
- Spatial consistency information can be potentially applied to help distinguish benign and adversarial instances against segmentation models.
- Strong adaptive attacker can hardly succeed when large randomness is incorporated into the model

Adversarial Frames In Videos

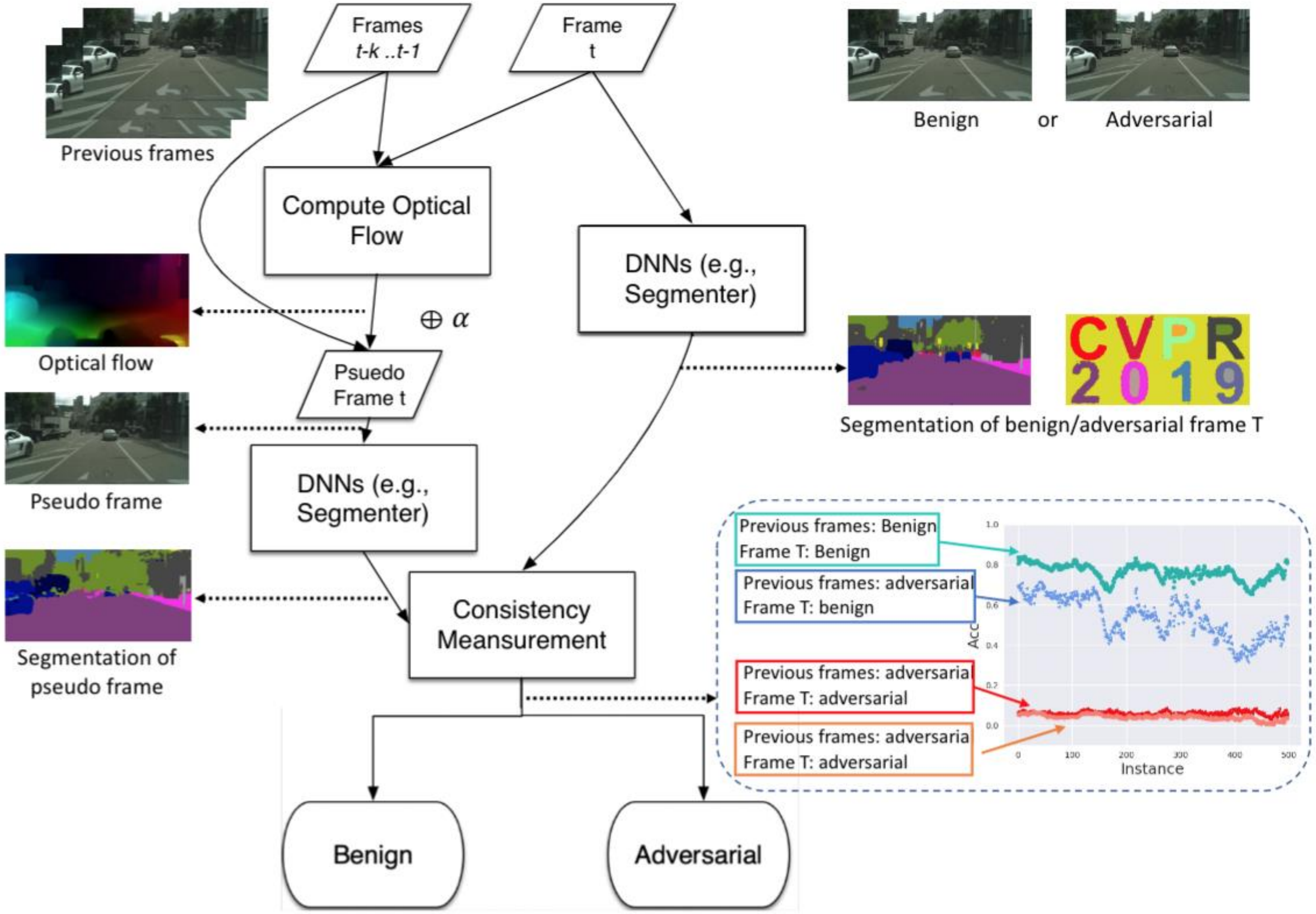
Attacks on segmentation

Attacks on pose estimation

Attacks on object detection

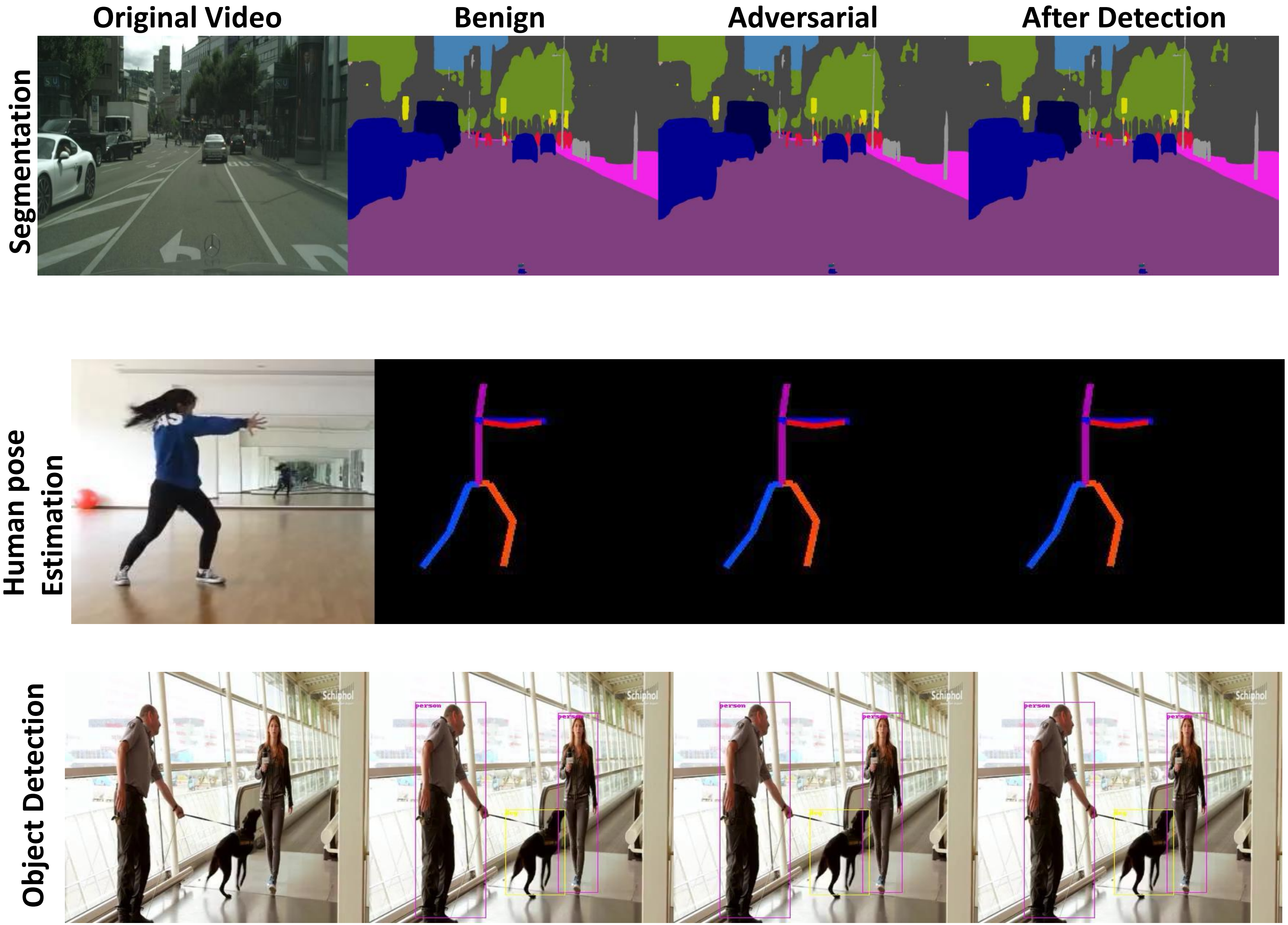


Defensing Adversarial behaviors in Videos – Temporal Dependency



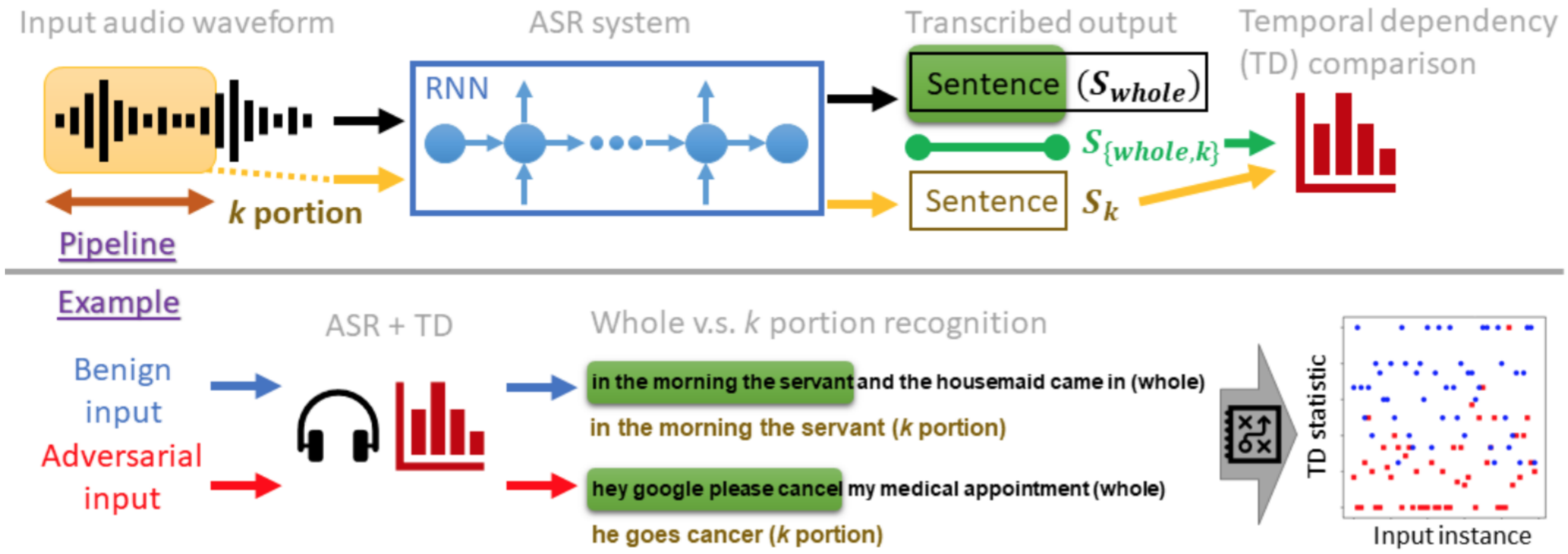
Task	Attack Method	Target	Previous Frames	Detection 1	Detection 3	Detection 5	Detection Adap 1	Detection Adap 3	Detection Adap 5
Semantic Segmentation	Houdini	CVPR	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	100%	100%	100%
		Remapping	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	100%	100%	100%
		Stripe	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	99%	100%	100%
	DAG	CVPR	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	100%	100%	100%
		Remapping	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	100%	100%	100%
		Stripe	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	100%	100%	100%
Human Pose Estimation	Houdini	shuffle	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	99%	100%	100%
		Transpose	Benign	100%	100%	100%	98%	100%	100%
			Adversarial	98%	99%	100%	98 %	99%	100%
Object Detection	DAG	all	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	98%	100%	100%
		person	Benign	99%	100%	100 %	100%	100%	100%
			Adversarial	97%	98%	100%	96 %	97%	100%

- The results show that choosing more random patches can improve detection rate while k=5 is enough to achieve AUC 100%
- The spatial consistency based detection is robust against strong adaptive attackers due to the randomness in patch selection


Original Video
Benign
Adversarial
After Detection
Segmentation
Human pose Estimation
Object Detection
Schiphol
person
dog

Temporal Consistency Based Analysis

- “Yanny” or “Laurel”? – adversarial audio



PeerNets: Exploiting Peer wisdom against adversarial attacks

- Design robust neural networks that are robust to adversarial attacks
- Defense: recover the ground truth instead of just tell adversarial instance apart
- Necessary step: design novel and advanced architectures built on new computational paradigms
- PeerNets:
    - Euclidean convolutions -> graph convolutions
    - Non-local forward propagation: Capture global structure induced by the data graph
    - Design a peer regularization layer

PeerNets: Exploiting Peer wisdom against adversarial attacks

- Peer Regularization layer
- For N images, each image will look for its K nearest neighbors based on cosine similarity
  - For each image, there is a $n \times d$ feature map

$$\tilde{\mathbf{x}}_p^i = \sum_{k=1}^{K} \alpha_{ij_k p q_k} \mathbf{x}_{q_k}^{j_k}, \qquad \alpha_{ij_k p q_k} = \frac{\mathrm{LeakyReLU}(\exp(a(\mathbf{x}_p^i, \mathbf{x}_{p_k}^{j_k})))}{\sum_{k'=1}^{K} \mathrm{LeakyReLU}(\exp(a(\mathbf{x}_p^i, \mathbf{x}_{p_{k'}}^{j_{k'}})))}$$

PeerNets: Exploiting Peer wisdom against adversarial attacks

- Randomized approximation
- Monte Carlo approximation
  - Select smaller batch and sample the nearest neighbor from each batch

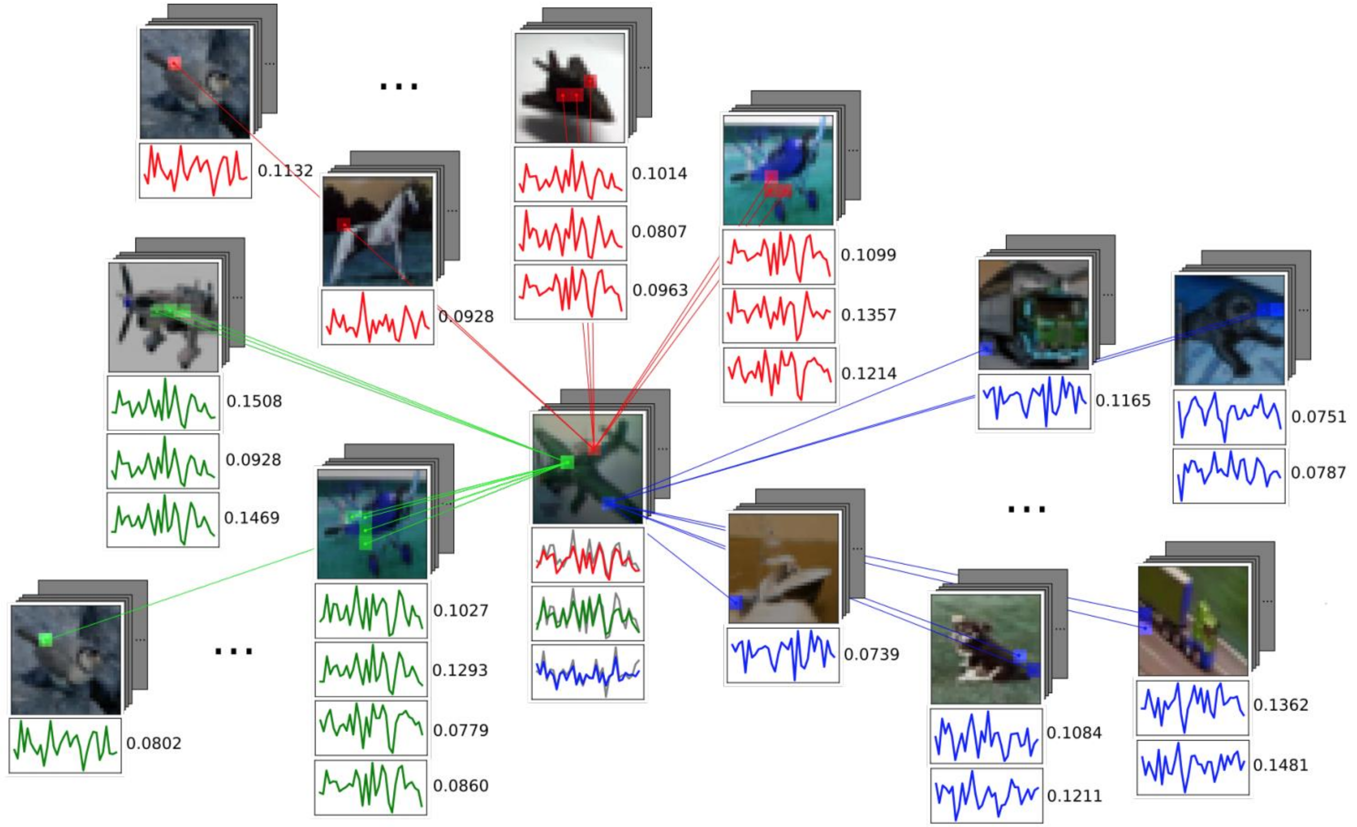
$$\{l_{m1}, \ldots, l_{mN}\} \subset \{1, \ldots, N'\}$$

$$\tilde{\mathbf{x}}_p^i = \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{ij_{mk}pq_{mk}} \mathbf{x}_{q_{mk}}^{j_{mk}}$$

- Other optimization method?

PeerNets: Exploiting Peer wisdom against adversarial attacks

- Select M = 1 during training and large M during inference
  - Limitations?

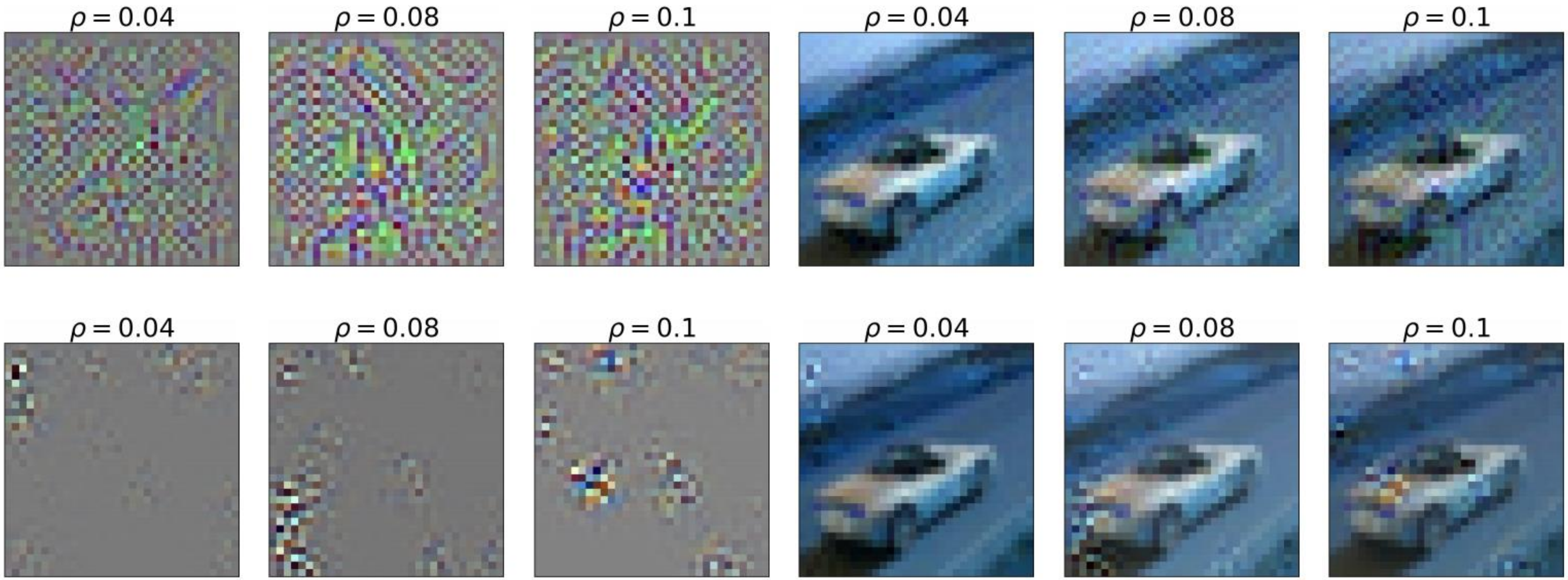


Results for PeerNets

Original

Reconstructed

Visualization of perturbation



Takeaways

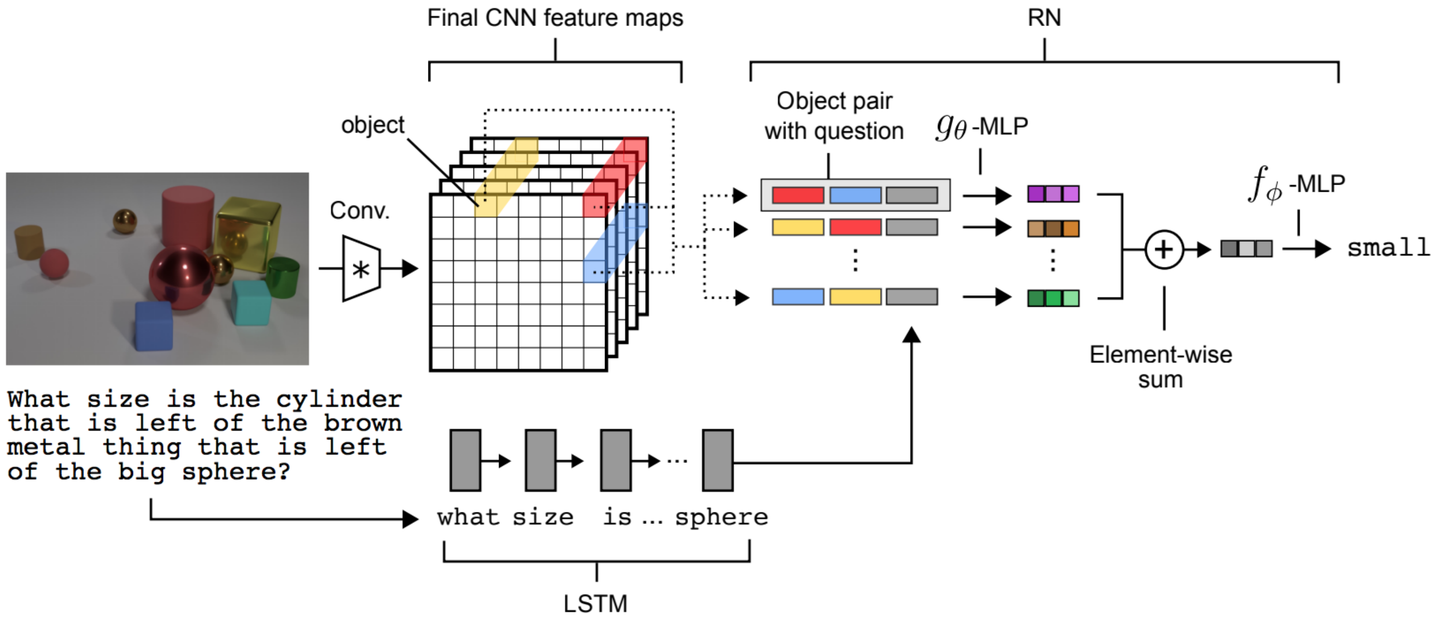
- Alternate Euclidean Graph convolution to harness information from peers can provide global information
- Can be added to any models as regularized layer –> good principle
- Not affect the benign accuracy -> important
- How to scale up?
- How to consider more peer images instead of pixels?
- Temporal information?

Similar reading

- Countering adversarial images using input transformations
  - Image quilting – nearest patches
  - Computationally expensive

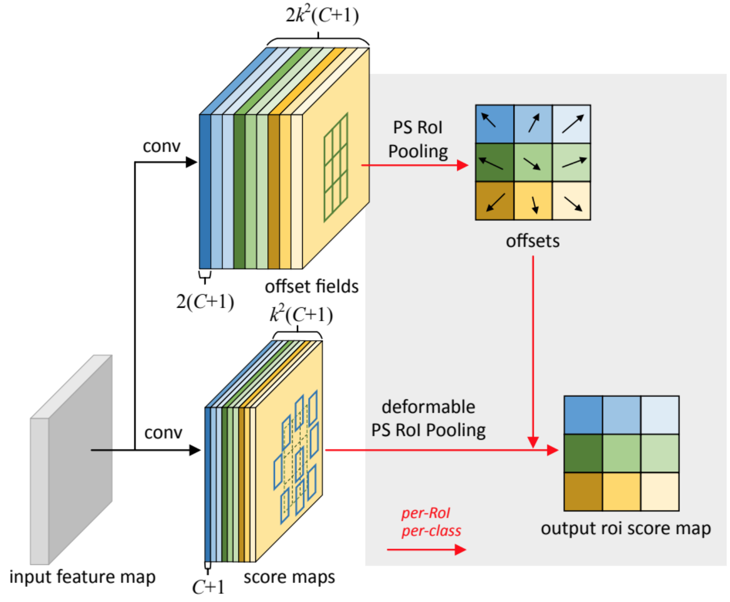
Interesting reading

- A simple neural network module for relational reasoning



Interesting reading

- Deformable Convolutional Networks
  - Deformable convolution and deformable RoI pooling
  - Augment the spatial sapling locations with additional offset which can be learned



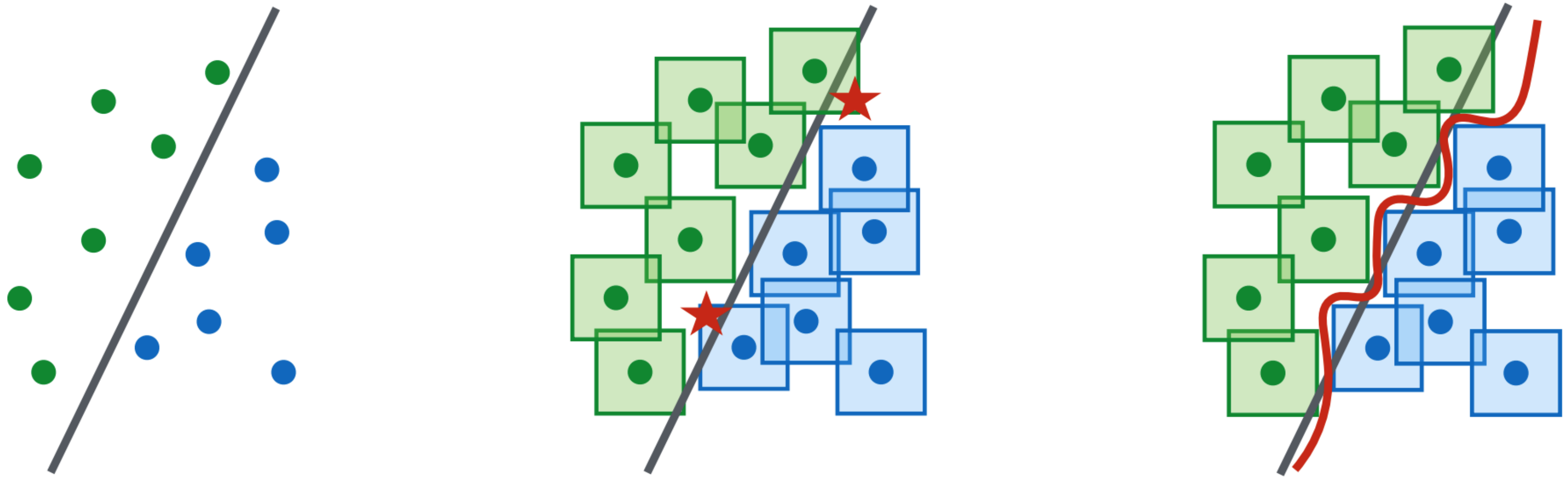
Towards Deep Learning Models Resistant to Adversarial Attacks

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in \mathcal{S}} L(\theta, x + \delta, y) \right]$$

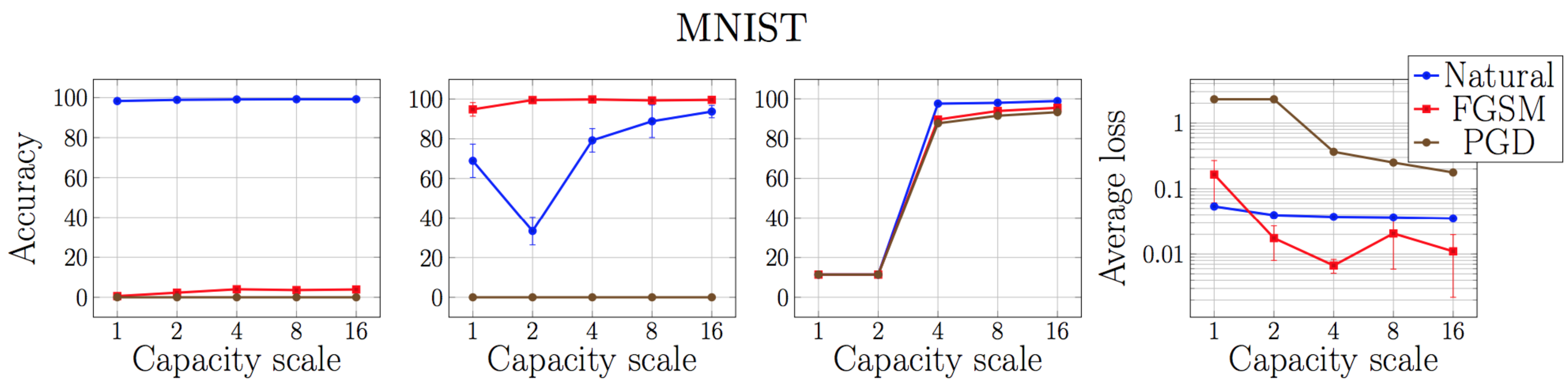
- Use a natural saddle point (min-max) formulation to capture the notion of security against adversarial attacks in a principled manner.
- The formulation casts both attacks and defenses into a common theoretical framework.
- Motivate projected gradient descent (PGD) as a universal “first-order adversary”.

Madry et al. Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2017.

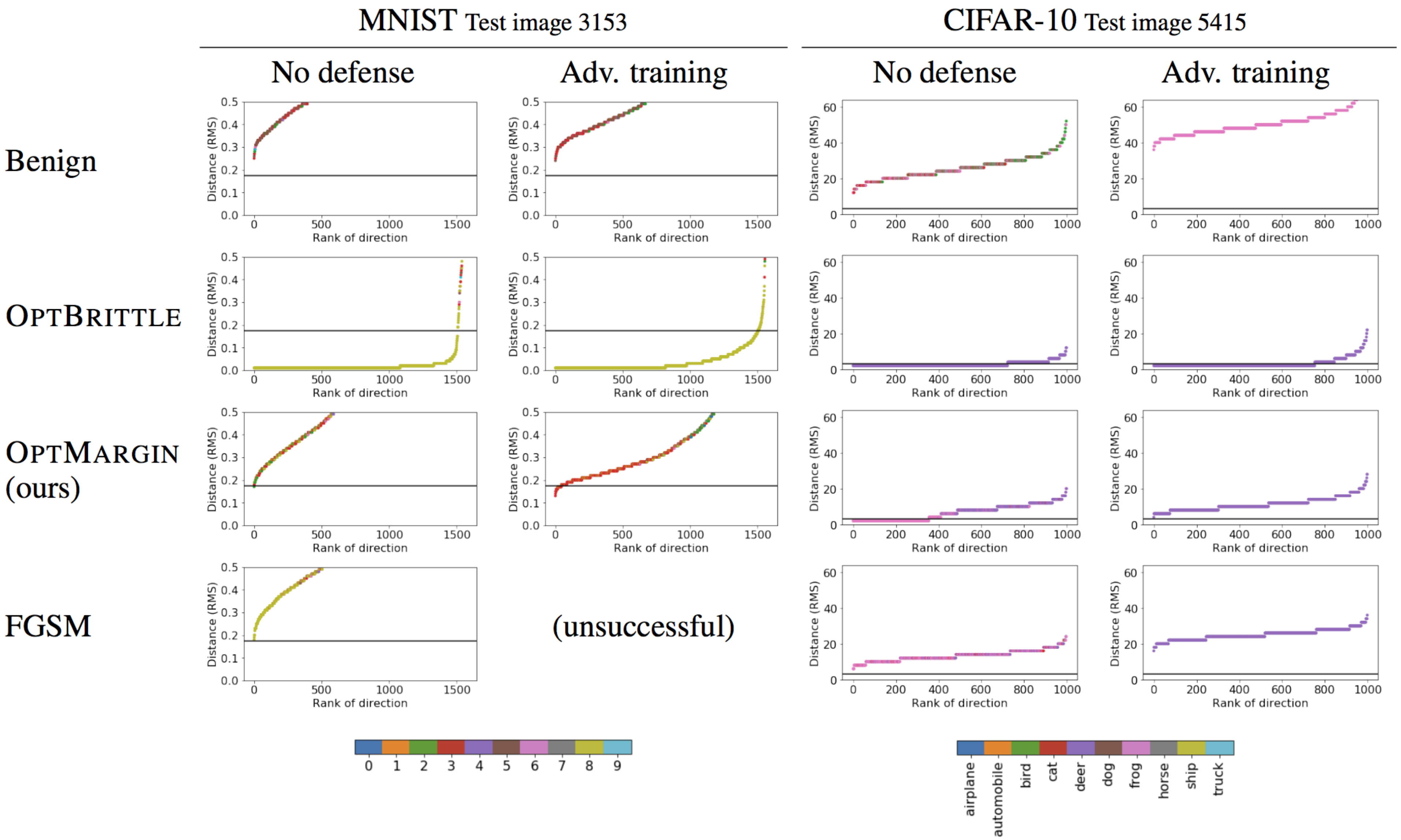
Model Capacity



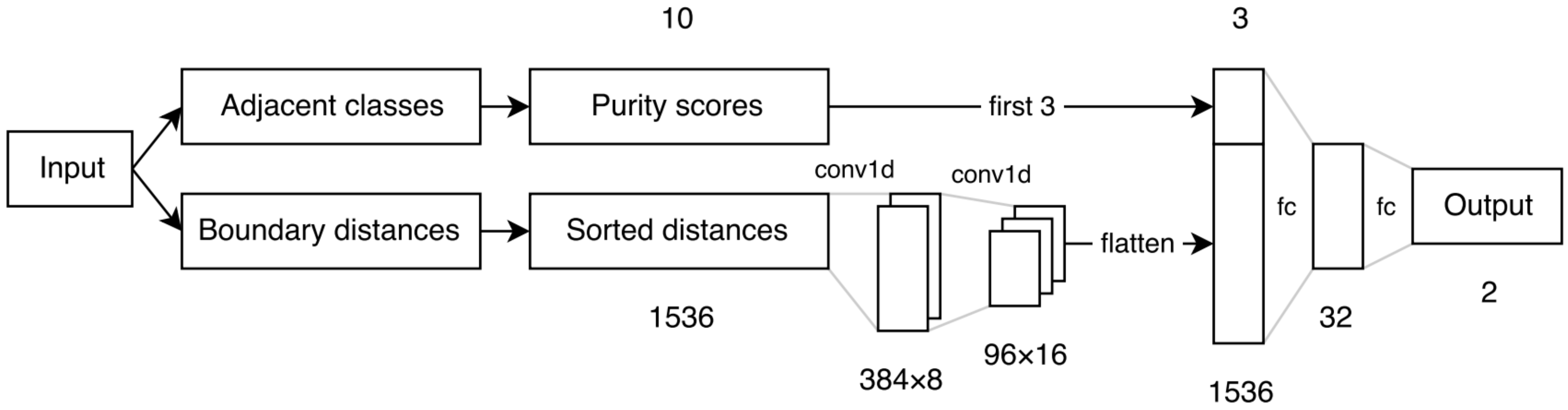
Towards Deep Learning Models Resistant to Adversarial Attacks



Decision Boundary Based Detection



Decision Boundary Analysis of Adversarial Examples



Training attack	False pos. Benign	False neg. OPTBRITTLE	False neg. OPTMARGIN	Accuracy Our approach	Accuracy Cao & Gong
MNIST, normal training				90.4%	10%
OPTBRITTLE	1.0%	1.0%	74.1%		
OPTMARGIN	**9.6%**	0.6%	7.2%		
MNIST, PGD adversarial training					
OPTBRITTLE	2.6%	2.0%	39.8%		
OPTMARGIN	10.3%	0.4%	14.5%		
CIFAR-10, normal training				96.4%	5%
OPTBRITTLE	5.3%	3.2%	56.8%		
OPTMARGIN	8.4%	7.4%	5.3%		
CIFAR-10, PGD adversarial training					
OPTBRITTLE	0.0%	2.4%	51.8%		
OPTMARGIN	**3.6%**	0.0%	1.2%		

Takeaways

- Decision boundaries of DNNs are important towards improving learning robustness
- Isolated islands in the data manifold would lead to harder detected/defensed adversarial behaviors