Indirect Local Attacks for Context-aware Semantic Segmentation Networks

European Conference of Computer Vision 2020

ID: 3995, Spotlight

Krishna Kanth Nakka and Mathieu Salzmann

Indirect Local Attack on Segmentation Networks

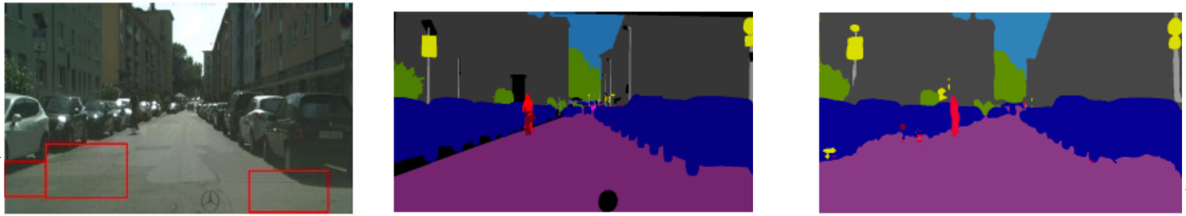
- **Contribution.** We expose the vulnerability of context-aware segmentation networks to indirect local attacks, where perturbation in static class region effects the prediction in dynamic class region**.**

- Experiment Setting.
  - Perturbation inside static class regions
    - ➢ road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky
  - Fooling the dynamic class regions
    - ➢ person, rider, car, truck, bus, train, motorcycle, bicycle
  - Targeted attack
    - ➢ Dynamic class regions are fooled to output the (spatially) nearest static class label (e.g. car -> road, bus -> road)
    - ➢ potentially creating a collision in autonomous driving scenario.

Overview: Indirect Local Attacks

We discover that modern *context-aware networks* are vulnerable to indirect local attacks.
Particularly, the location of perturbation and fooling is **different**.

Imperceptible perturbations inside few static regions shown as red boxes



(a) Adversarial image (b) Ground Truth (c) FCN

(No context)

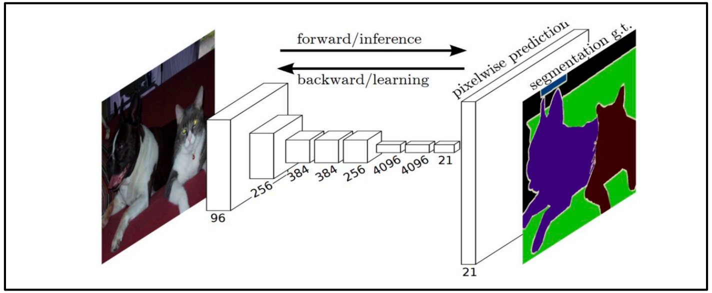
FCN is robust to attacks

(d) PSPNet (e) PSANet (f) DANet

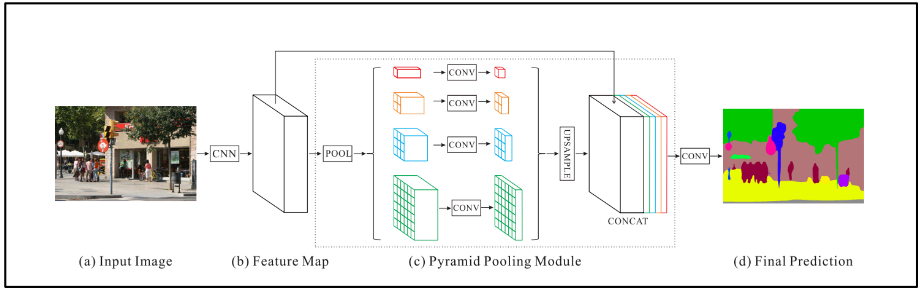
(context by pooling) ***(context by point-wise spatial attention)*** ***(context by spatial & channel attention)***

Dynamic regions belonging to car, pedestrians far away from perturbed area are effected in modern networks (PSANet, PSPNet, DANet) that use surrounding context

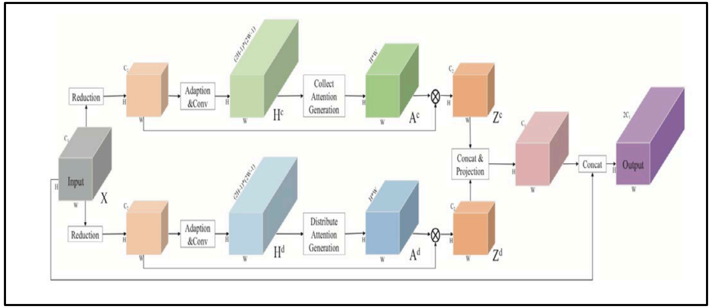
Context in Semantic Segmentation Networks



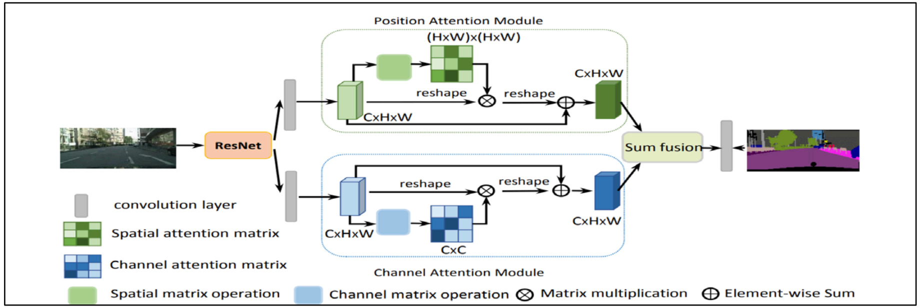
1. Vanilla FCN



2. PSPNet: Context by spatial pyramid pooling



3. PSANet: Context by pointwise spatial attention



4. DANet: Context by spatial & channel attention

Indirect Attacks

- Image-dependent indirect attacks
  - perturbation location – predetermined
  - perturbation location – optimized to be within few patches

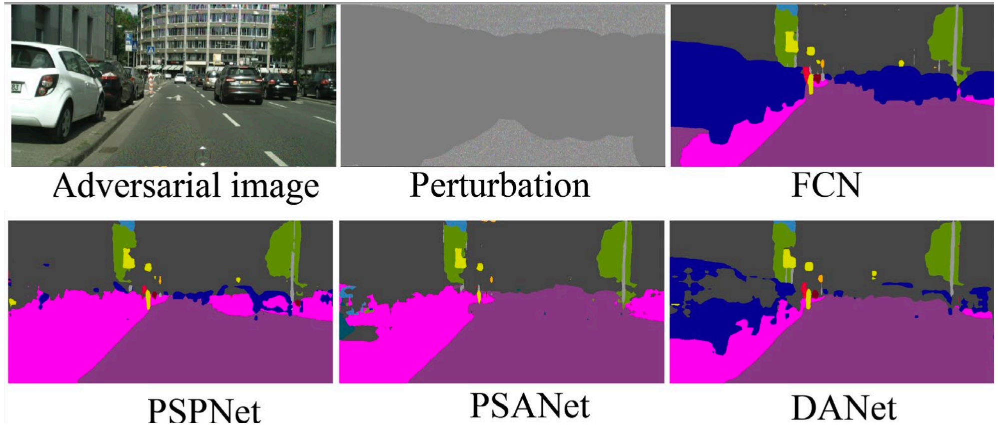
- Image-independent indirect attacks
  - universal indirect attacks

- Metrics:

  mIoU_u - mIoU computed b/w adversarial and normal sample predictions
  ASR_t - percentage of pixels that were predicted as the target label

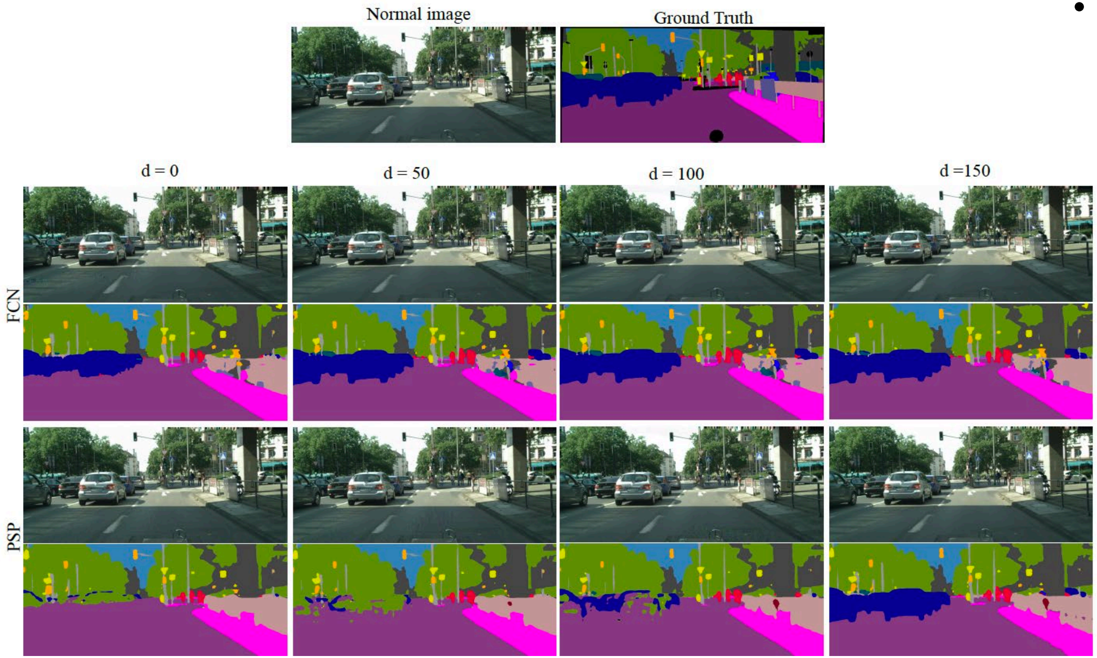
1. Indirect Attack

- Perturbation location inside static pixel regions
    - **predetermined**
    - parametric distance d from dynamic class objects



Impact of indirect attacks by perturbing static class pixels that are at least d =100 pixels away from any dynamic class for a 512 x 1024 input image

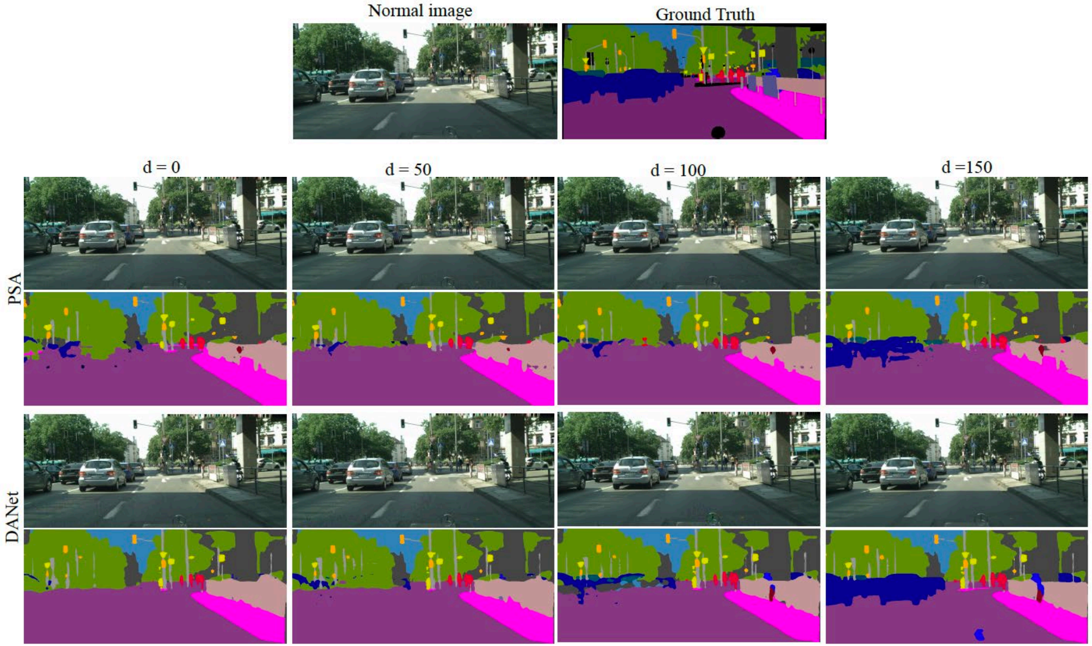
1. Indirect Attack



Impact of indirect attacks by perturbing static class pixels that are at least d pixels away from any dynamic class for a 512 x 1024 input image

- Perturbation location inside static class regions
  - Predetermined
  - not optimized
  - parametric distance d from dynamic class objects

1. Indirect Attack



Impact of indirect attacks by perturbing static class pixels that are at least d pixels away from any dynamic class for a 512 x 1024 input image

- Perturbation location inside static class regions
  - Predetermined
  - not optimized
  - parametric distance d from dynamic class objects

Experiments: Indirect Attack

Network	$d = 0$	$d = 50$	$d = 100$	d = 150
FCN [29]	0.11 / <u>64%</u>	0.77 / <u>2.0%</u>	0.98 / <u>0%</u>	1.00 / <u>0.0%</u>
PSPNet [53]	0.00 / **90%**	0.14 / 73%	0.24 / 60%	0.55 / 23%
PSANet [54]	0.00 / **90%**	0.11 / 71%	0.13 / **65%**	0.29 / **47%**
DANet [12]	0.00 / **90%**	0.13 / **81%**	0.48 / 43%	0.80 / 10%
DRN [50]	0.02 / 86%	0.38 / 22%	0.73 / 3%	0.94 / 1.0%

$\mathrm{mIoU}_u/\mathrm{ASR}_t$

(a) ℓ_∞ attack

Impact of local attacks by perturbing pixels that are at least *d* pixels away from any dynamic class.

2. Adaptive Indirect Local Attack

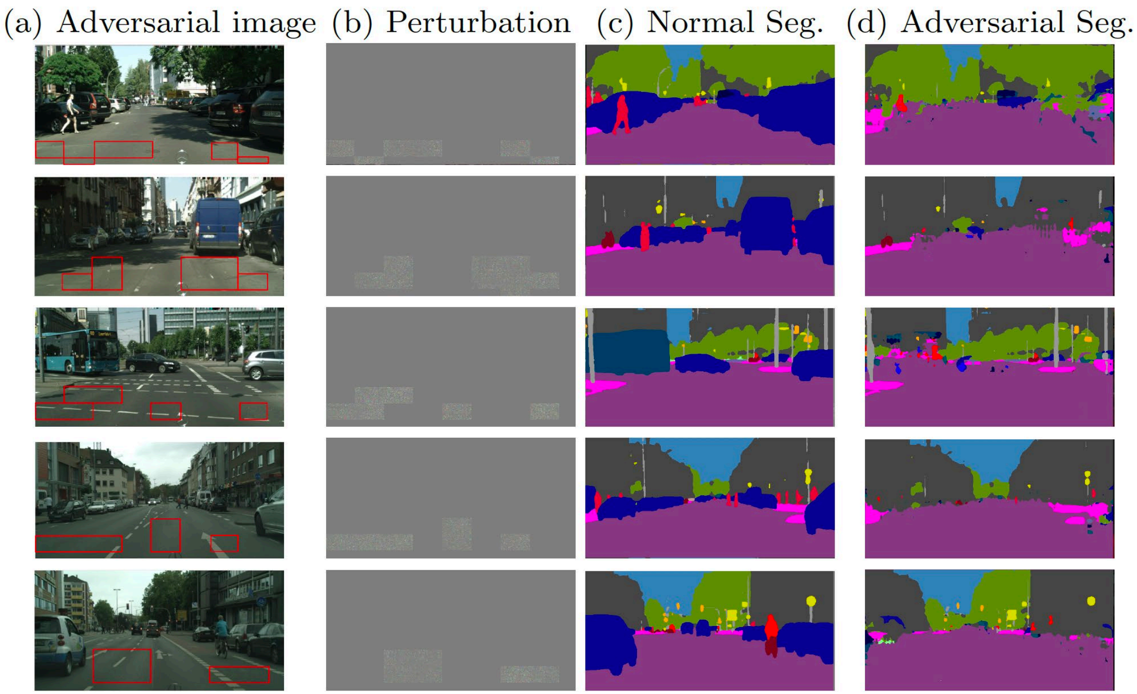
Optimally find the best locations to perturb

$$\delta^* = \arg\min_{\delta} \ \lambda_2 \sum_{t=1}^{T} \|\mathbf{M}_t \odot \delta\|_2 + \lambda_1 \|\delta\|_2^2 + J_t(\mathbf{X}, \mathbf{M}, \mathbf{F}, \delta, f, \mathbf{y}^{pred}, \mathbf{y}^t)$$

T	:	number of patches	$\mathbf{X}$	:	input image
δ	:	perturbation	y^{pred}	:	predicted label map
$\mathbf{M}$	:	perturbation mask	y^t	:	targeted label map
$\mathbf{F}$	:	fooling mask			

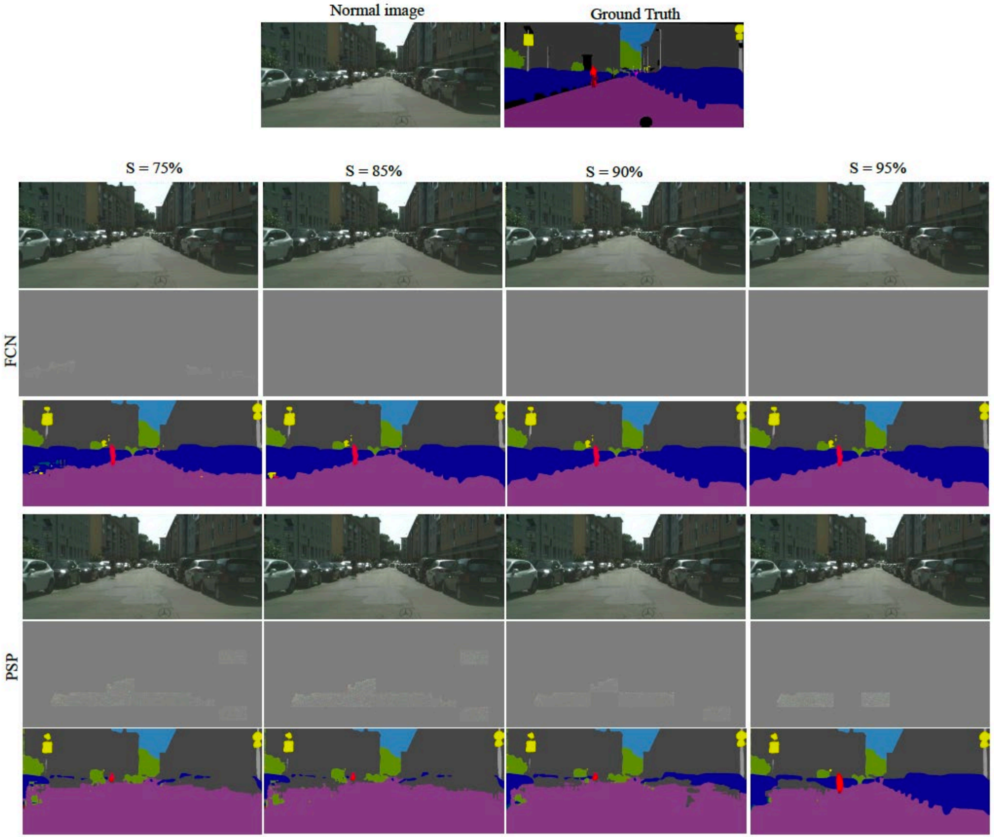
2. Adaptive Indirect Attack on PSANet

- Perturbation location
  - Confined to few patches in static regions
  - Optimized by group sparsity prior at patch level



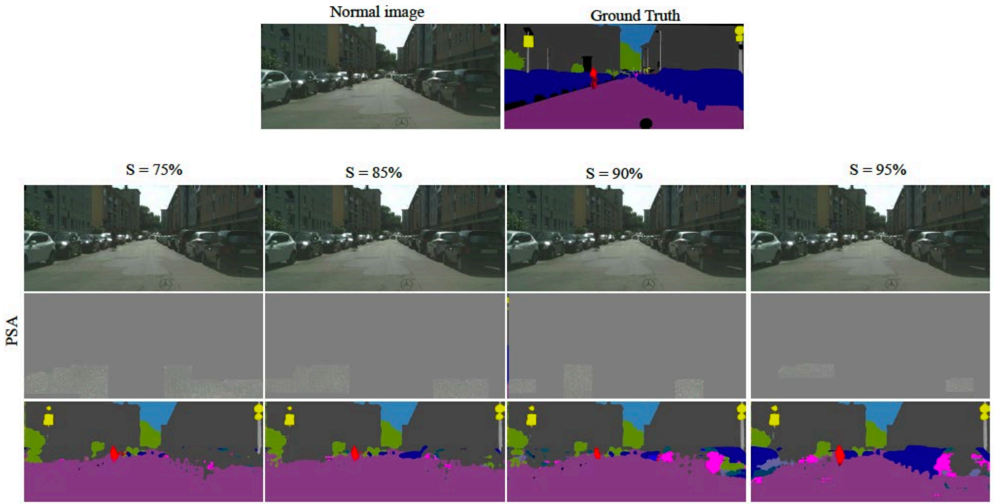
Cityscapes

2. Adaptive Indirect Attack



- Perturbation location
  - ➢ Confined to few patches in static regions
  - ➢ Optimized by group sparsity prior at patch level
- Sparsity
  - ➢ percentage of pixels that are not perturbed relative to the initial perturbation mask

2. Adaptive Indirect Attack

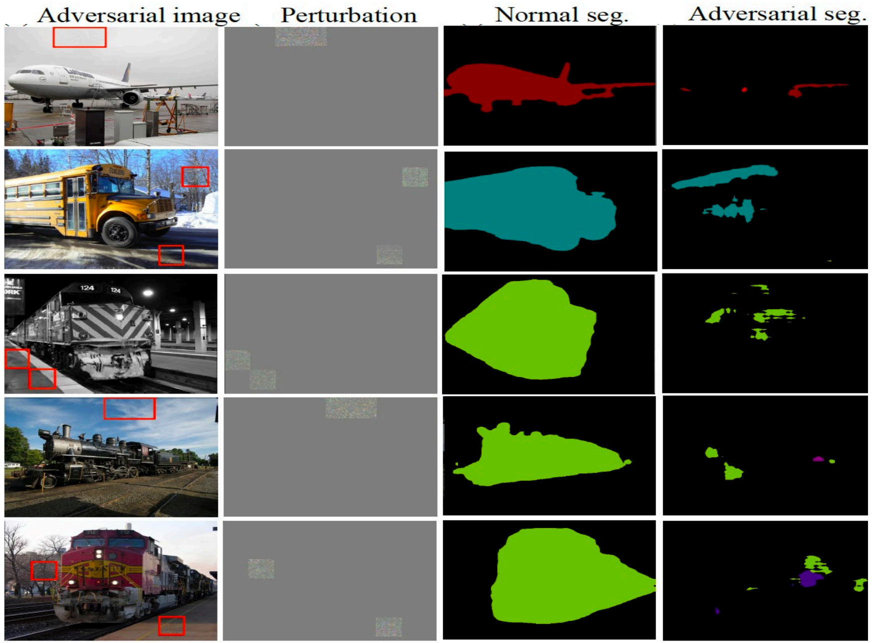


Cityscapes

- Perturbation location
  - ➢ Confined to few patches in static regions
  - ➢ Optimized by group sparsity prior at patch level
- Sparsity
  - ➢ percentage of pixels that are not perturbed relative to the initial perturbation mask

2. Adaptive Indirect Attack on PSANet

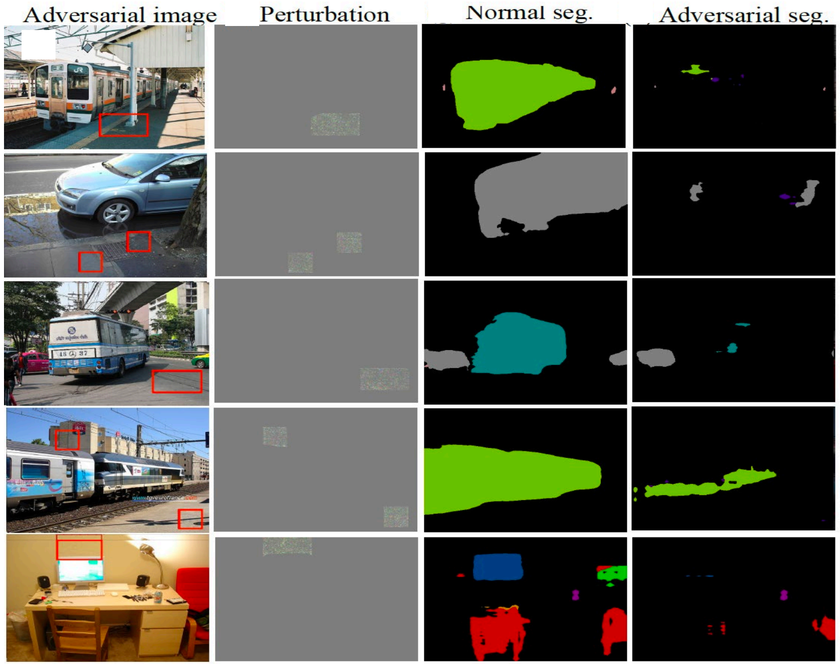
- Perturbation location
  - Confined to few patches in static regions
  - Optimized by group sparsity prior at patch level



PASCAL VOC

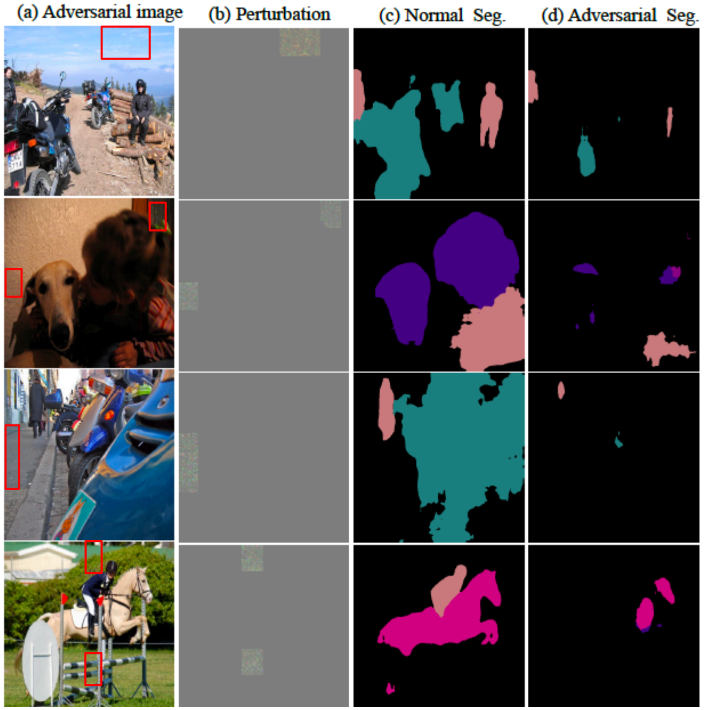
2. Adaptive Indirect Attack on PSANet

- Perturbation location
  - Confined to few patches in static regions
  - Optimized by group sparsity prior at patch level



PASCAL VOC

2. Adaptive Indirect Attack on PSANet



PASCAL VOC

- Perturbation location
  - Confined to few patches in static regions
  - Optimized by group sparsity prior at patch level

2. Adaptive Indirect Attack on PSANet

PASCAL VOC

- Perturbation location
  - ➢ Confined to few patches in static regions
  - ➢ Optimized by group sparsity prior at patch level

Experiments: Adaptive Indirect Attack

Network	$S = 75\%$	$S = 85\%$	$S = 90\%$	$S = 95\%$
FCN [29]	0.52 / 12%	0.66 / 6%	0.73 / 4%	0.84 / 1.0%
PSPNet [53]	0.19 / 70%	0.31 / 54%	0.41 / 42%	0.53 / 21%
PSANet [54]	0.10 / **78%**	0.16 / **71%**	0.20 / **64%**	0.35 / **44%**
DANet [12]	0.30 / 64%	0.52 / 43%	0.64 / 30%	0.71 / 21%
DRN [50]	0.42 / 23%	0.55 / 13%	0.63 / 9%	0.77 / 4.5%

$\text{mIoU}_u/\text{ASR}_t$

(a) Cityscapes

Network	$S = 75\%$	$S = 85\%$	$S = 90\%$	$S = 95\%$
FCN [29]	0.50 / 32%	0.59 / 27%	0.66 / 22%	0.80 / 12%
PSANet [54]	0.28 / **68%**	0.21 / **77%**	0.20 / **80%**	0.30 / **69%**

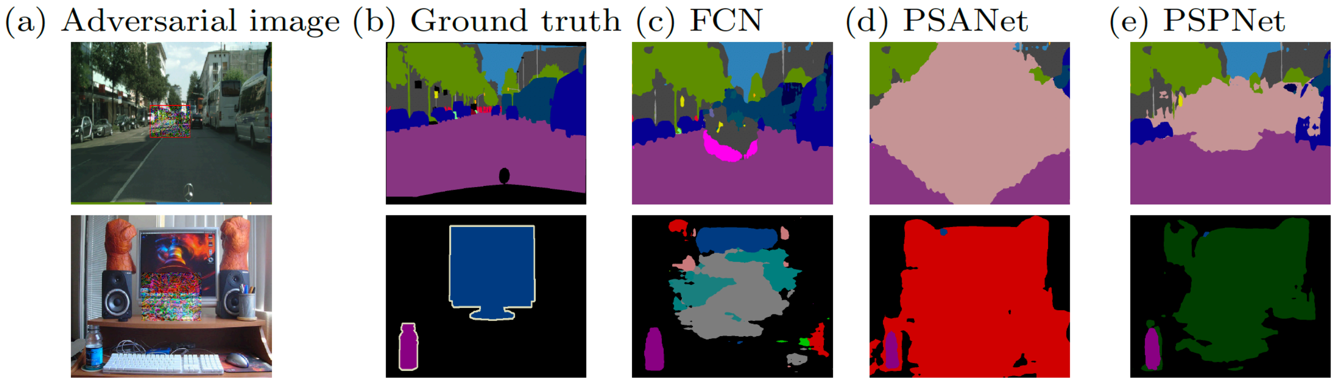
$\text{mIoU}_u/\text{ASR}_t$

(b) PASCAL VOC

Performance of adaptive indirect local attacks for a given sparsity level

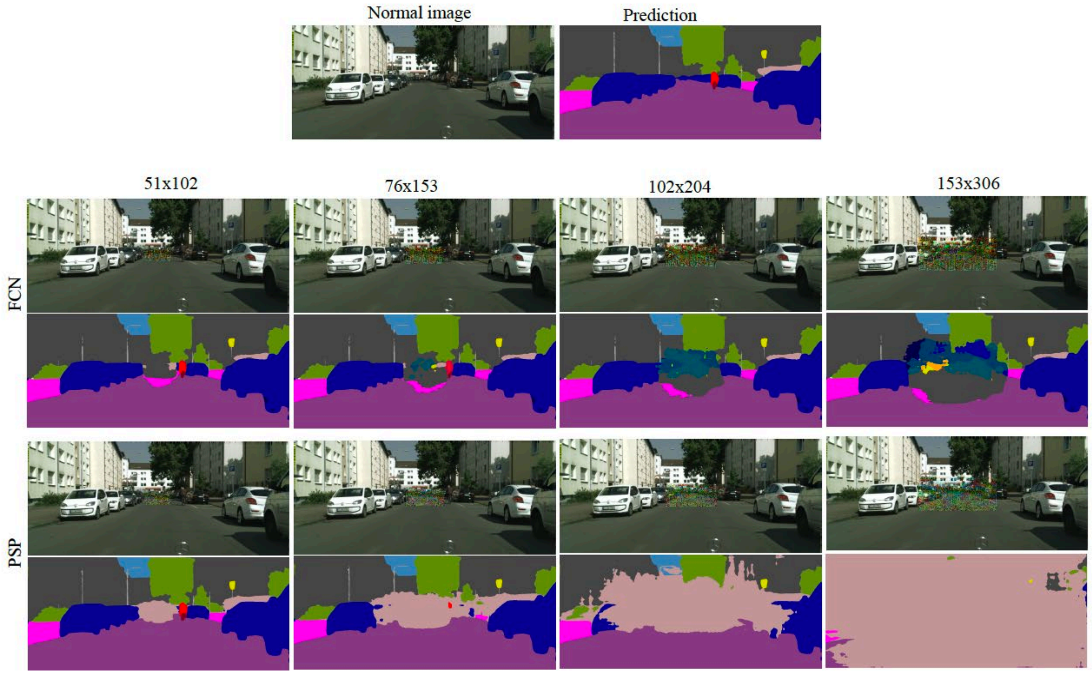
3. Universal Attack

- Image-independent
- Perturbation location
    - confined to a single patch at the center
- Untargeted attack to fool entire image



Universal local attacks on Cityscapes and PASCAL VOC using a single fixed size patch

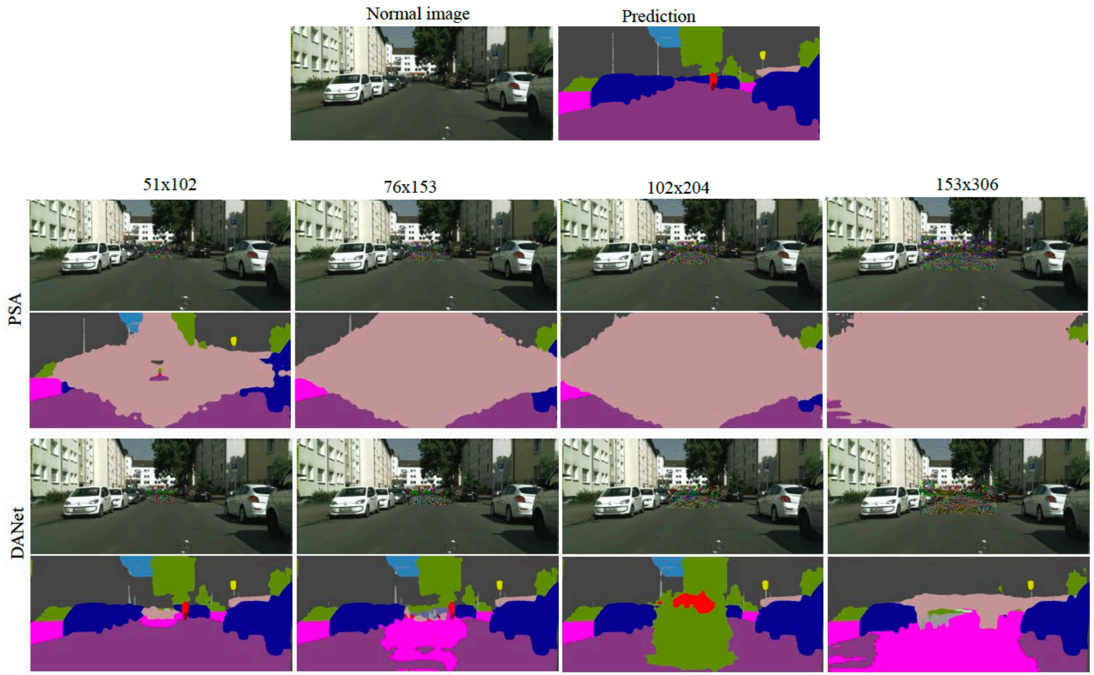
3. Universal Attack



Universal local attacks on Cityscapes using a different sizes of patch

- Image-independent
- Perturbation location
  - ➢ confined to a single patch at the center
- Untargeted attack to fool entire image

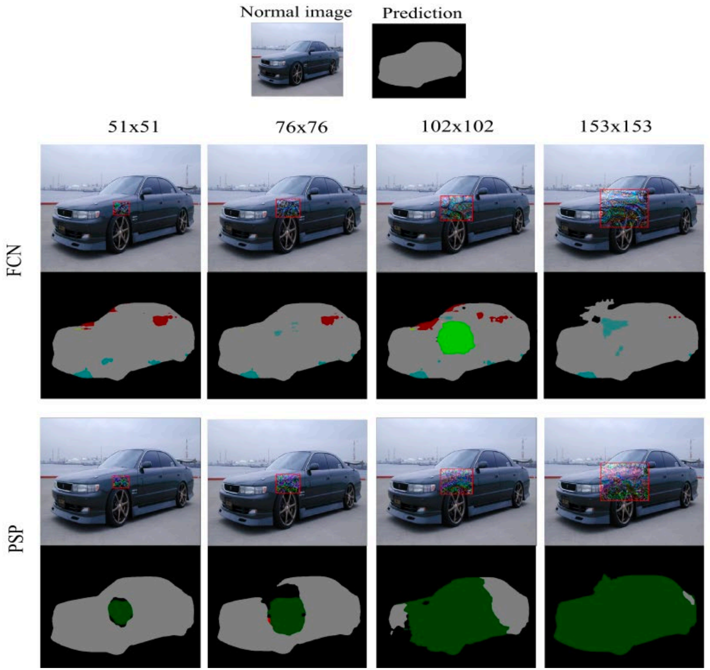
3. Universal Attack



Universal local attacks on Cityscapes using a different sizes of patch

- Image-independent
- Perturbation location
    - ➢ confined to a single patch at the center
- Untargeted attack to fool entire image

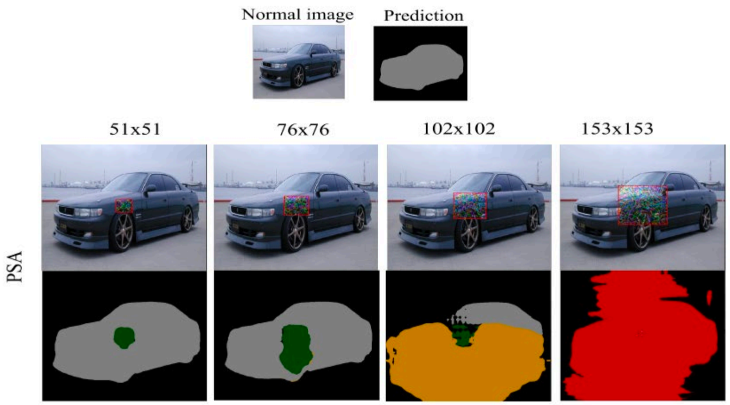
3. Universal Attack



Universal local attacks on PASCAL VOC using a single fixed size patch

- Image-independent
- Perturbation location
  - confined to a single patch at the center
- Untargeted attack to fool entire image

3. Universal Attack



Universal local attacks on PASCAL VOC using a single fixed size patch

- Image-independent
- Perturbation location
    - ➢ confined to a single patch at the center
- Untargeted attack to fool entire image

Experiments: Universal Attack

Network	$51 \times 102(\mathbf{1.0\%})$	$76 \times 157(\mathbf{2.3\%})$	$102 \times 204(\mathbf{4.0\%})$	$153 \times 306(\mathbf{9.0\%})$
FCN [29]	0.85 / 2.0%	0.78 / 4.0%	0.73 / 9.0%	0.58 / 18%
PSPNet [53]	0.79 / 3.0%	0.63 / 11%	0.44 / 27%	0.08 / 83%
PSANet [54]	0.41 / **37%**	0.22 / **60%**	0.14 / 70%	**0.10** / **90%**
DANet [12]	0.79 / 4.0%	0.71 / 10%	0.65 / 15%	0.40 / 42%
DRN [50]	0.82 / 3.0%	0.78 / 8.0%	0.71 / 14%	0.55 / 28%

$\text{mIoU}_u/\text{ASR}_u$

(a) Cityscapes

Impact of universal local attacks by perturbing patch of size h x w (area%) for 512 x 1024 input image

Experiments: Attack Detection

- We detect the region of fooling by computing Mahalanobis distance between feature and nearest class-conditional distribution at every spatial location j

$$C(\mathbf{X}_j^\ell) = \max_{c\in[1,C]} -\left(\mathbf{X}_j^\ell - \mu_c^\ell\right)^\top \mathbf{\Sigma}_\ell^{-1}\left(\mathbf{X}_j^\ell - \mu_c^\ell\right)$$

$\mathbf{X}_j^\ell$ - feature at location j and layer l

μ_c^ℓ - Class-specific mean at layer l

$\mathbf{\Sigma}^\ell$ - covariance at layer l

Experiments: Attack Detection

Networks	Perturbation region	Fooling region	ℓ_∞ / ℓ_2 norm	Mis. pixels %	Global AUROC SC [48] / Re-Syn [25] / **Ours**	Local AUROC **Ours**
FCN [29]	Global	Full	0.10 / 17.60	90%	**1.00** / **1.00** / 0.94	0.90
	UP	Full	0.30 / 37.60	4%	0.71 / 0.63 / **1.00**	0.94
	FS	Dyn	0.07 / 2.58	13%	0.57 / 0.71 / **1.00**	0.87
	AP	Dyn	0.14 / 3.11	1.7%	0.51 / 0.65 / **0.87**	0.89
PSPNet [53]	Global	Full	0.06 / 10.74	83%	0.90 / **1.00** / 0.99	0.85
	UP	Full	0.30 / 38.43	11%	0.66 / 0.70 / **1.00**	0.96
	FS	Dyn	0.03 / 1.78	14%	0.57 / 0.75 / **0.90**	0.87
	AP	Dyn	0.11 / 5.25	11%	0.57 / 0.75 / **0.90**	0.82
PSANet [54]	Global	Full	0.05 / 8.26	92%	0.90 / 1.00 / 1.00	0.67
	UP	Full	0.30 / 38.6	60%	0.65 / 1.00 / 1.00	0.98
	FS	Dyn	0.02 / 1.14	12%	0.61 / 0.76 / **1.00**	0.92
	AP	Dyn	0.10 / 5.10	10%	0.50 / 0.82 / **1.00**	0.94
DANet [12]	Global	Full	0.06 / 12.55	82%	0.89 / 1.00 / 1.00	0.68
	UP	Full	0.30 / 37.20	10%	0.67 / 0.63 / **0.92**	0.89
	FS	Dyn	0.05 / 1.94	13%	0.57 / 0.69 / **0.94**	0.88
	AP	Dyn	0.14 / 6.12	43%	0.59 / 0.68 / **0.98**	0.82

Attack detection on Cityscapes with different perturbation settings

Global – full image perturbations
UP - universal patch perturbations.

FS – full static region perturbations
AP – adaptive attack perturbations

Summary

- We show the vulnerability of modern context-aware networks to various indirect attacks
- We propose adaptive indirect attack based on group sparsity
- We evaluate the impact of context to universal fixed-size patch attacks
- We propose pixel-level detection of fooling regions based on Mahalanobis distance