On the Robustness of Semantic Segmentation Models to Adversarial Attacks

Anurag Arnab, Ondrej Miksik, Philip H.S. Torr

UNIVERSITY OF OXFORD

1. Introduction

Adversarial examples are arguably the greatest challenge affecting DNNs. No effective defence exists yet [1].

- We investigate how robust modern DNN-based semantic segmentation models are to adversarial examples.
- We show connections between architectural features of segmentation networks and recently proposed defences [2,3,4].
- We also show that the "conventional wisdom" derived from image classification does not always hold on different tasks and large-scale datasets.

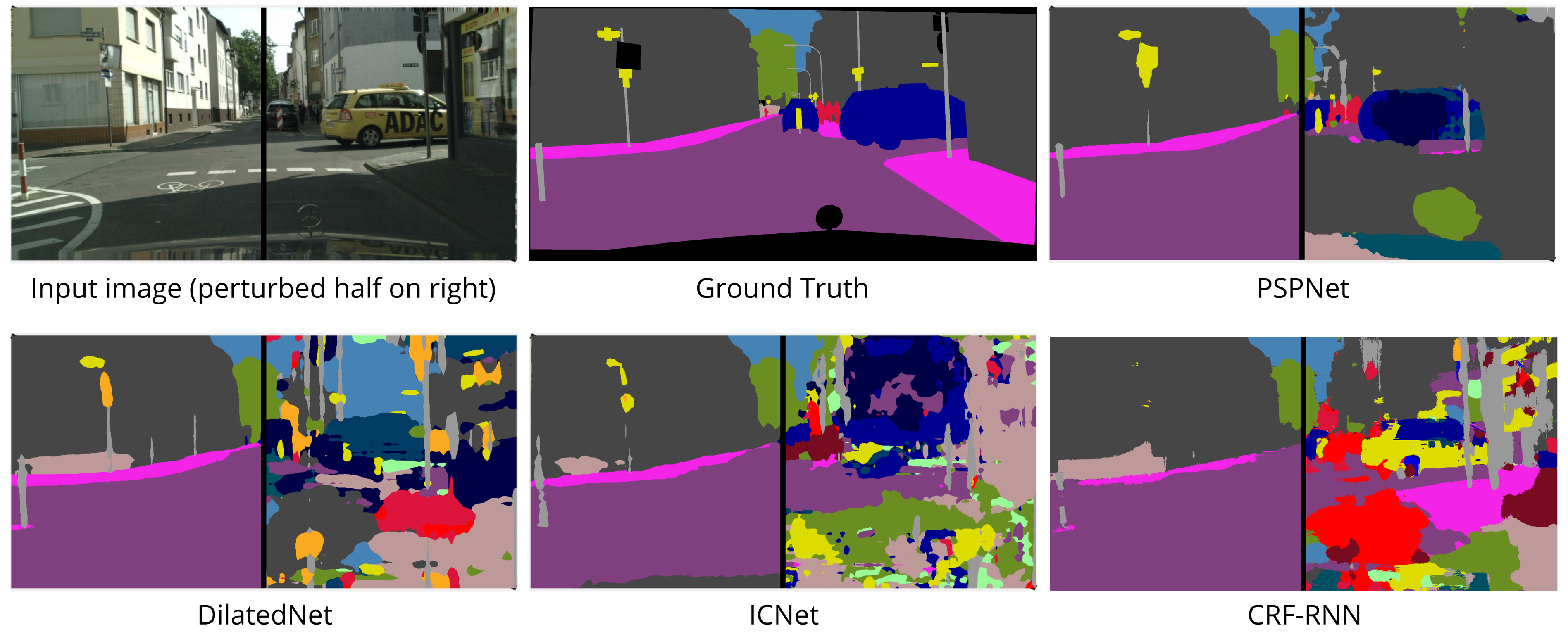


Fig. 1. Adversarial example created with an imperceptible ℓ_∞ norm of 4. All networks are severely affected, but to different degrees.

2. Experimental set-up

We evaluate state-of-art models (Fig. 2) on the Cityscapes and Pascal VOC datasets. We use the *IoU Ratio* metric to account for varying clean accuracy.

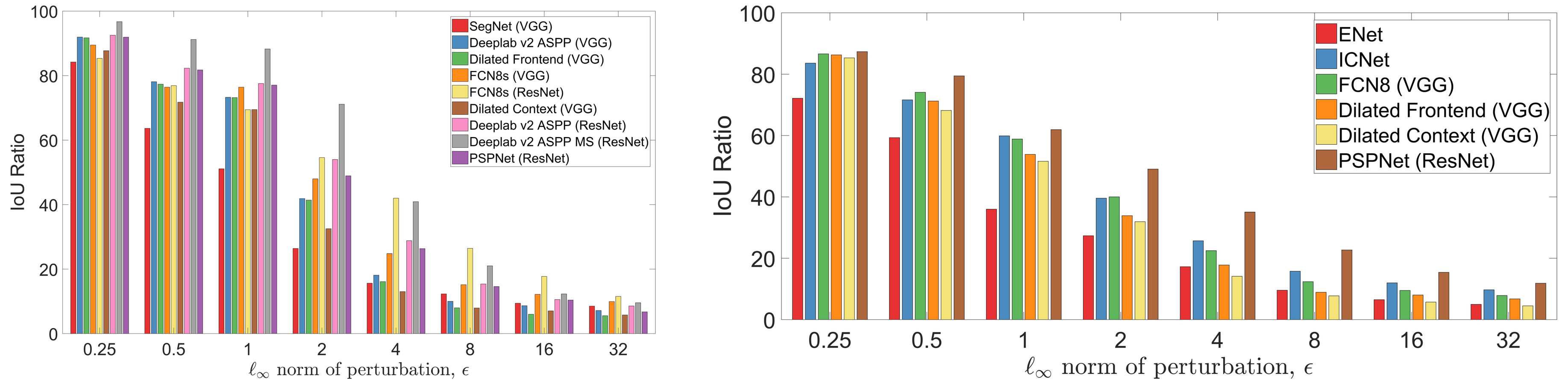
We used variants of the FGSM attack for varying ℓ_∞ norm constraints.

FGSM: $\mathbf{x}^{adv} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} L(f(\mathbf{x};\theta), y))$.

Iterative FGSM ll: $\mathbf{x}_{t+1}^{adv} = \text{clip}(\mathbf{x}_t^{adv} - \alpha \cdot \text{sign}(\nabla_{\mathbf{x}_t^{adv}} L(f(\mathbf{x}_t^{adv};\theta), y_{ll})), \epsilon)$.

3. Robustness of various DNN architectures

Figure 2 evaluates several state-of-art architectures on Cityscapes and VOC.



a) Pascal VOC b) Cityscapes

Fig. 2. Robustness of various models on VOC (a) and Cityscapes (b). Models are ordered according to clean accuracy.

- Models with residual connections are inherently more robust than chain-like VGG-based networks
- This holds also for lightweight models (E-Net and ICNet), contrary to [5, 6].
- Accuracy on clean inputs and robustness is correlated, though the most accurate (PSPNet) is not the most robust model (DeepLab v2 MS).
- Perturbations that do not change integral RGB values degraded all models.

4.1 Multiscale Processing

- Multiscale processing (Deeplab v2) increases robustness, and white-box attacks on such networks produce more transferable black-box perturbations.
- CNNs are not invariant to scale (and many other transformations). As such, predictions on rescaled adversarial inputs change to become less malignant. Same effect when the network is trained with or without multiscale averaging.

Table 1. Transferability of perturbations from different scales of Deeplab v2 $(\epsilon = 8)$

Network	FGSM				Iterative FGSM ll			
	50%	75%	100%	Multiscale	50%	75%	100%	Multiscale
Deeplab v2 50%	37.3	70.5	84.8	**60.3**	18.0	92.0	96.9	**20.0**
Deeplab v2 75%	85.5	39.7	62.2	**50.8**	99.5	17.9	89.9	**20.4**
Deeplab v2 100%	93.6	57.9	37.7	**37.2**	100.0	79.0	15.5	**16.8**
Deeplab v2 Mutiscale	83.7	**57.6**	62.3	53.1	99.6	**90.2**	91.9	21.5
Deeplab v2 100% (VGG)	94.3	70.6	66.9	**66.5**	98.9	88.4	86.3	**80.9**
FCN8 (VGG)	94.7	67.2	65.8	**65.4**	98.4	85.2	84.9	**78.5**
FCN8 (ResNet)	94.0	66.3	63.5	**63.1**	99.4	82.6	80.3	**74.1**

4.2 Other Input Transformations

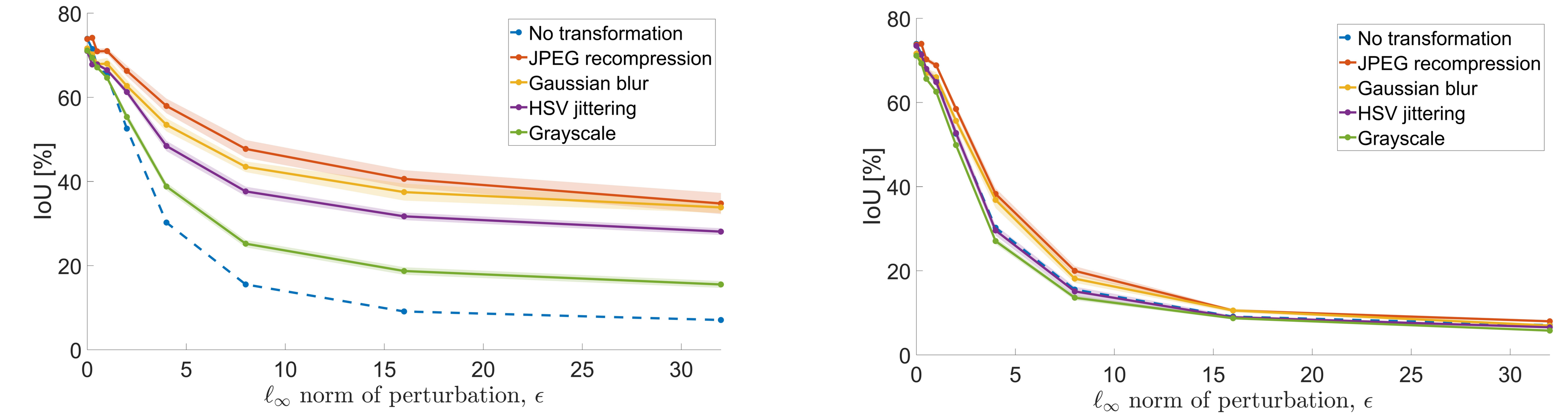


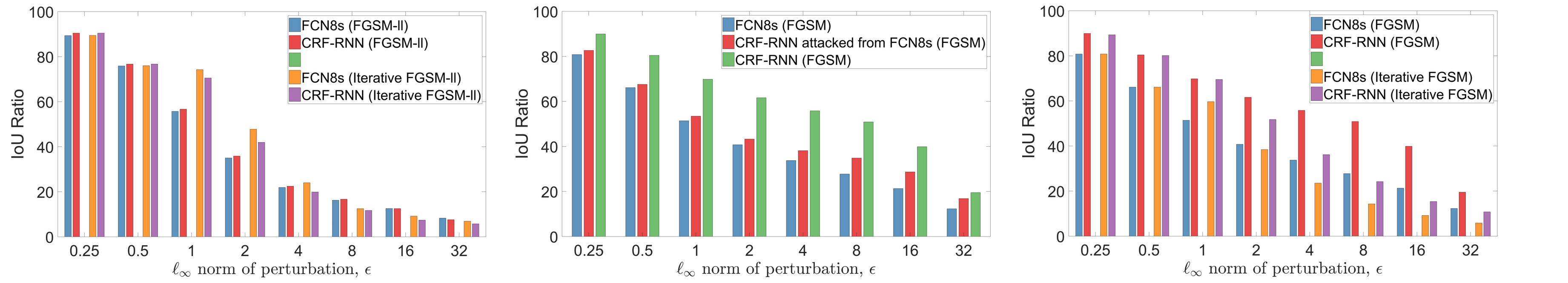
Fig. 3. Randomised input transformations only confer robustness when the attack is oblivious to it (left). Otherwise, their benefits are marginal (right).

- Many other transformations (besides scale) that CNNs are not invariant to.
- Performed JPEG recompression, Gaussian blur, HSV jittering and Grayscale conversion with randomised parameters. In all cases, randomised input transformations markedly increased robustness (Fig. 3).
- Easily subverted if we include knowledge of transformation into the attack

$$\mathbf{x}_{t+1}^{\text{adv}} = \text{clip}\left(\mathbf{x}_t^{\text{adv}} - \alpha \cdot \text{sign}(\mathbb{E}_{t\sim\mathcal{T}} \nabla_{\mathbf{x}_t^{\text{adv}}} L(f(t(\mathbf{x}_t^{\text{adv}});\theta), y_{ll}), \epsilon)\right).$$

- Many proposed defences based on input transformations (i.e. [2, 3] among others) were not evaluated correctly (and do not apply Kerckhkoff's principle [7]).
- Corroborates findings that producing physical adversarial examples is difficult [8] - an object may undergo a myriad of transformations before camera capture.

5. Conditional Random Fields (CRFs)



a) Untargeted attack b) Targeted attack c) Black-box attack

Fig. 4. Mean-field inference of CRFs produces confident estimates which mask gradients. As such, it is only robust to untargeted attacks.

- CRFs are often used to enforce smoothness and other consistency priors.
- Mean-field inference naturally performs gradient masking.
- CRFs are thus more robust to white-box attacks, but vulnerable to targeted attacks and transfer, black-box attacks as shown in Fig. 4.

[1] A Athalye *et al.* Obfuscated gradients give a false sense of security: Circumventing defences to adversarial examples. *ICML* 2018.
[2] N Papernot *et al.* Distillation as a defense to adversarial perturbations against deep neural networks. *IEEE SSP* 2016
[3] C Xie *et al.* Mitigating adversarial effects through randomisation. *ICLR* 2018.
[4] C Guo *et al.* Countering adversarial images using input transformations. *ICLR* 2018.
[5] A Madry *et al.* Towards deep learning models resistant to adversarial attacks. *ICLR* 2018.
[6] A Kurakin *et al.* Adversarial machine learning at scale. *ICLR* 2017.
[7] A Kerckhoffs. La cryptographie militaire. *Journal des sciences militaire* 1883.
[8] J Lu *et al.* Standard detectors aren't fooled by physical adversarial stop signs. *arXiv* 2017.

www.robots.ox.ac.uk/~aarnab | aarnab@robots.ox.ac.uk