On the Robustness of Semantic Segmentation Models to Adversarial Attacks

Anurag Arnab Ondrej Miksik Philip H.S. Torr
University of Oxford
{anurag.arnab, ondrej.miksik, philip.torr}@eng.ox.ac.uk

Organized by: Xi Fang

Date: 09/09/2020

Motivation

- Deep Neural Networks have demonstrated exceptional performance on most recognition tasks.
- However, they have also been shown to be vulnerable to adversarial examples.
- Numerous strategies have been proposed to train DNNs to be more robust to adversarial examples
- These defenses are not universal; they have frequently been found to be vulnerable to other types of attacks
- Adversarial examples have not been extensively analysed beyond standard image classification models, and often on small datasets such as MNIST or CIFAR-10 by the time

Contribution

- They present the first rigorous evaluation of the robustness of semantic segmentation models to adversarial attacks
- They focus on semantic **segmentation**, since it is a significantly more complex task than image classification. Based on classification architecture, extended by additional components
- They show that adversarial examples are less effective when processed at different scales.
- They examine other input transformations which neural networks are not invariant to and show that they are markedly more robust to transformed adversarial examples.

Adversarial Examples

- Adversarial perturbations cause a classifier to change its original prediction, when added to the original input x.

$$\arg\min \quad \|\mathbf{r}\|_2 \qquad \text{subject to} \quad f(\mathbf{x}+\mathbf{r};\theta) = y_t,$$

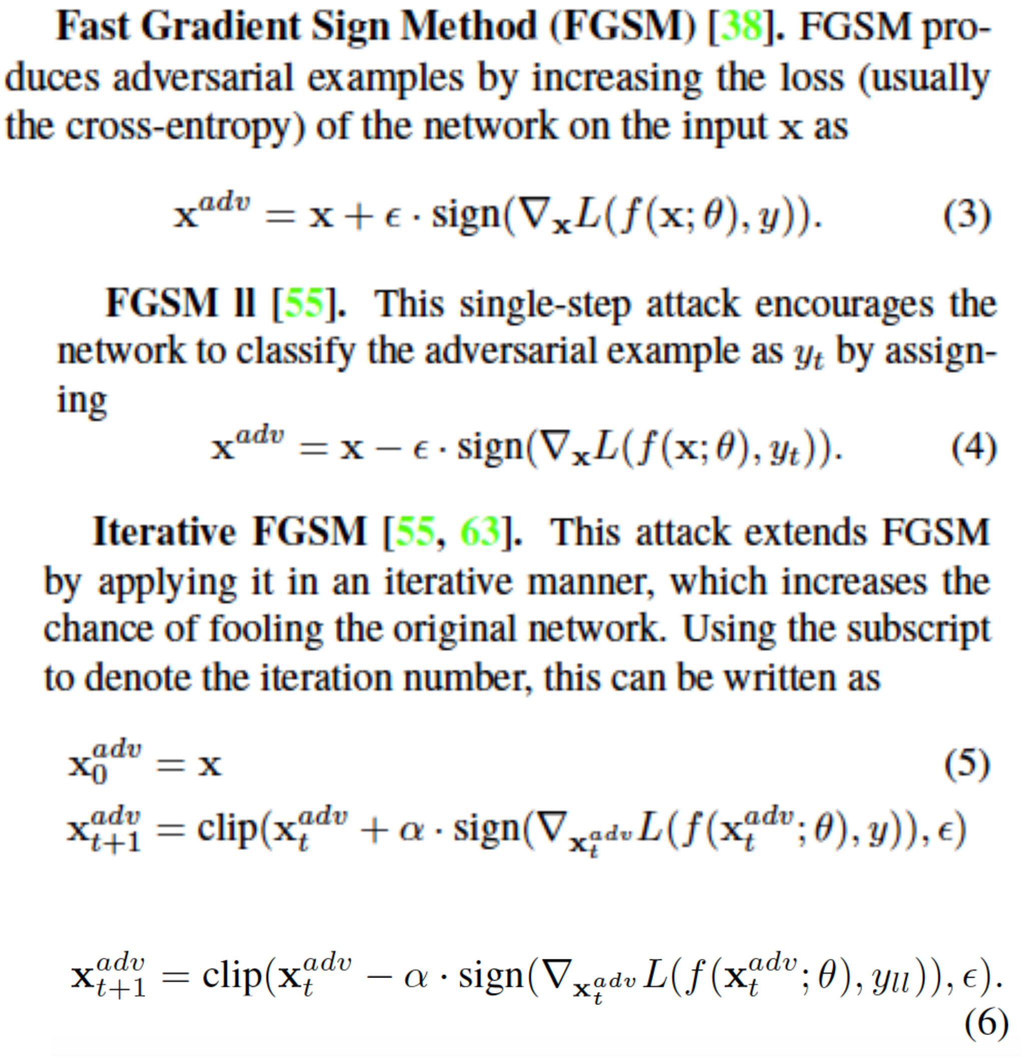
$$\mathbf{x}^{adv} = \mathbf{x} + \mathbf{r}.$$

- added an additional term to the objective based on the loss function used to train the network

$$\underset{\mathbf{r}}{\arg\min} \quad c\|\mathbf{r}\|_2 + L(f(\mathbf{x}+\mathbf{r};\theta), y_t).$$

Adversarial Samples

- Method
  - FGSM
  - FGSM II
  - Iterative FGSM
  - Iterative FGSM II



Fast Gradient Sign Method (FGSM) [38]. FGSM produces adversarial examples by increasing the loss (usually the cross-entropy) of the network on the input x as

$$\mathbf{x}^{adv} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} L(f(\mathbf{x};\theta), y)). \quad (3)$$

FGSM ll [55]. This single-step attack encourages the network to classify the adversarial example as y_t by assigning

$$\mathbf{x}^{adv} = \mathbf{x} - \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} L(f(\mathbf{x};\theta), y_t)). \quad (4)$$

Iterative FGSM [55, 63]. This attack extends FGSM by applying it in an iterative manner, which increases the chance of fooling the original network. Using the subscript to denote the iteration number, this can be written as

$$\mathbf{x}_0^{adv} = \mathbf{x} \quad (5)$$
$$\mathbf{x}_{t+1}^{adv} = \text{clip}(\mathbf{x}_t^{adv} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}_t^{adv}} L(f(\mathbf{x}_t^{adv};\theta), y)), \epsilon)$$

$$\mathbf{x}_{t+1}^{adv} = \text{clip}(\mathbf{x}_t^{adv} - \alpha \cdot \text{sign}(\nabla_{\mathbf{x}_t^{adv}} L(f(\mathbf{x}_t^{adv};\theta), y_{ll})), \epsilon). \quad (6)$$

Adversarial Defenses and Evaluation

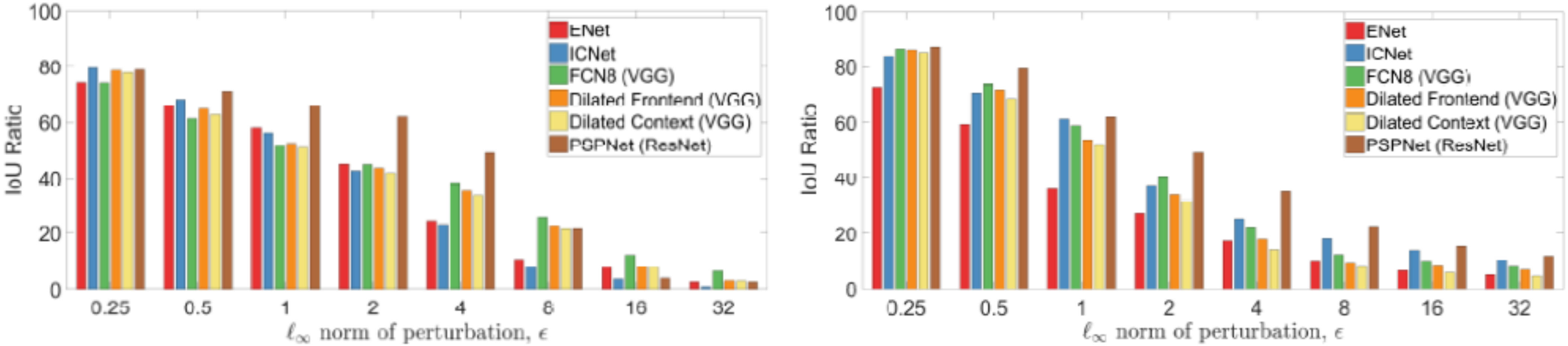
- No effective defense to all adversarial attacks exist. This motivates them to study the properties of state-of-the-art segmentation networks and how they affect robustness to various adversarial attacks.
- Moreover, our evaluation is carried out on two large-scale datasets instead of only ImageNet.
- The conclusions from the evaluations may thus aid future efforts to develop defenses to adversarial attacks that preserve predictive accuracy.

Experiment

- Datasets:
  - Pascal VOC consists of internet images labelled with 21 different classes.
  - Cityscapes consists of road-scenes captured from car-mounted cameras and has 19 classes.
- Models:
  - VGG, ResNet; ENet, ICNet
  - PSPNet, DeepLab, Segnet, E-Net,CRF, DilatedNet and FCN
- Adversarial attacks
  - FGSM, FGSM II, iterative FGSM, iterative FGSM II
- Evaluation metric
  - The Intersection over Union (IoU) is the primary metric used in evaluating semantic segmentation. However, as the accuracy of each model varies, IoU Ratio is used.

Findings

- The robustness of different architectures



(a) Untargeted FGSM attack

(b) Targeted Iterative FGSM ll. attack

Figure 3: Adversarial robustness of state-of-the-art models on the Cityscapes dataset. Contrary to Madry *et al.* [63], we observe that lightweight networks such as E-Net [74] and ICNet [91] are often about as robust as Dilated-Net [90] (341× more parameters than E-Net). Dilated-Net without its "Context" module is slightly more robust than the full network. As with the VOC dataset, ResNet (PSPNet) architectures are more robust than VGG (Dilated-Net and FCN8). Curiously, the FGSM attack is more effective than Iterative FGSM ll which computes adversarial examples from a larger search space.

Findings

- The unexpected effectiveness of single step methods on Cityscapes
- Imperceptible perturbations
  - With $\varrho = 0.25$, the perturbation is so small that the RGB values of the image pixels (assuming integers $\in [0, 255]$) are usually unchanged.

Findings

- Multiscale Processing and Transferability of Adversarial Examples
  - Robustness conferred by randomised input transformations

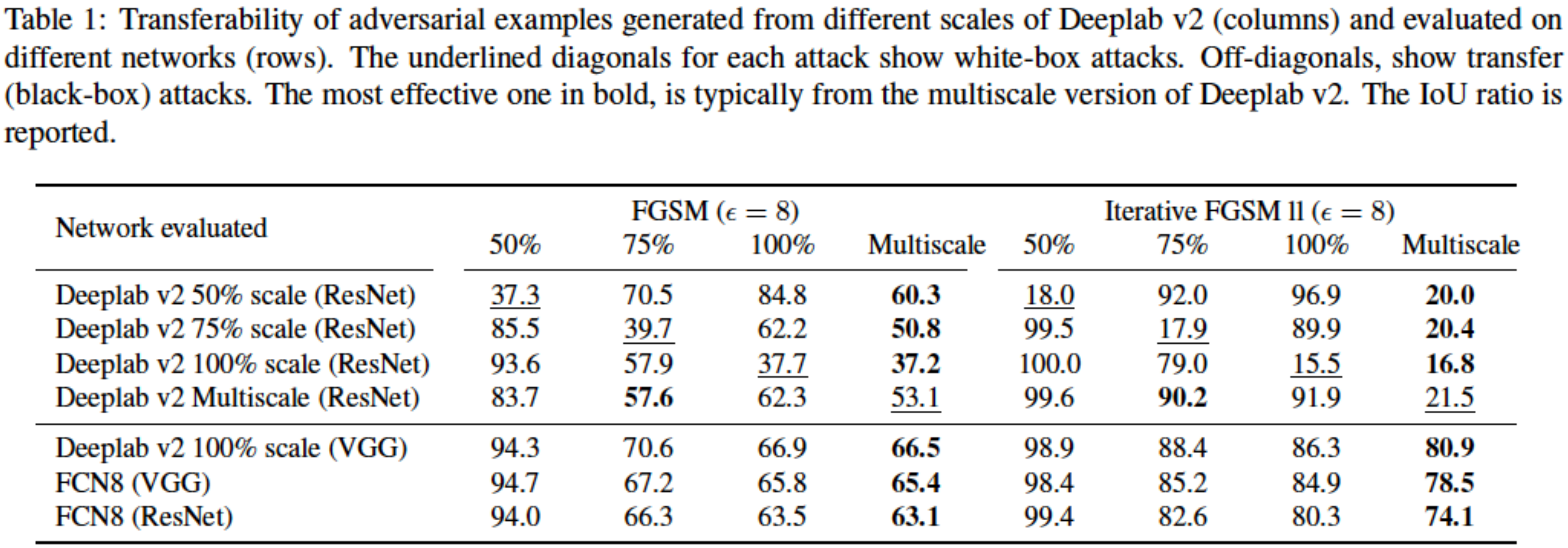
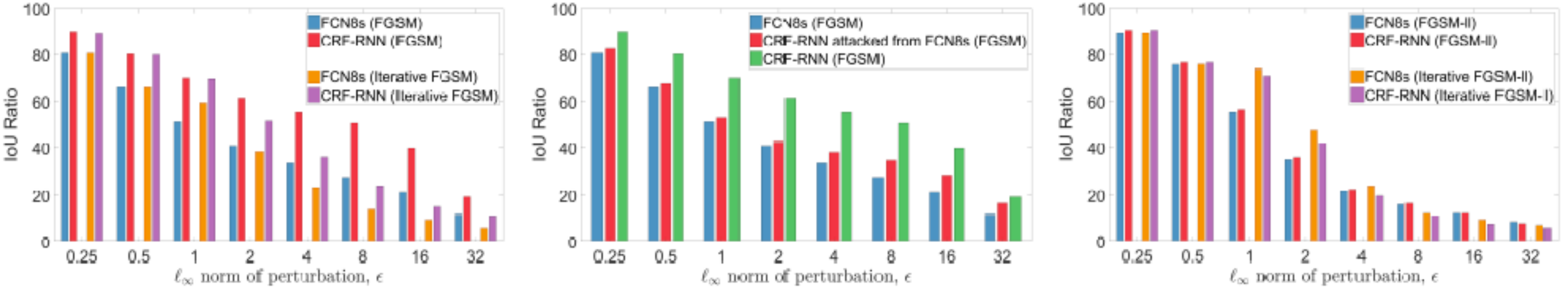


Table 1: Transferability of adversarial examples generated from different scales of Deeplab v2 (columns) and evaluated on different networks (rows). The underlined diagonals for each attack show white-box attacks. Off-diagonals, show transfer (black-box) attacks. The most effective one in bold, is typically from the multiscale version of Deeplab v2. The IoU ratio is reported.

Network evaluated	FGSM ($\epsilon = 8$)				Iterative FGSM ll ($\epsilon = 8$)			
	50%	75%	100%	Multiscale	50%	75%	100%	Multiscale
Deeplab v2 50% scale (ResNet)	37.3	70.5	84.8	**60.3**	18.0	92.0	96.9	**20.0**
Deeplab v2 75% scale (ResNet)	85.5	39.7	62.2	**50.8**	99.5	17.9	89.9	**20.4**
Deeplab v2 100% scale (ResNet)	93.6	57.9	37.7	**37.2**	100.0	79.0	15.5	**16.8**
Deeplab v2 Multiscale (ResNet)	83.7	**57.6**	62.3	53.1	99.6	**90.2**	91.9	21.5
Deeplab v2 100% scale (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**

Findings

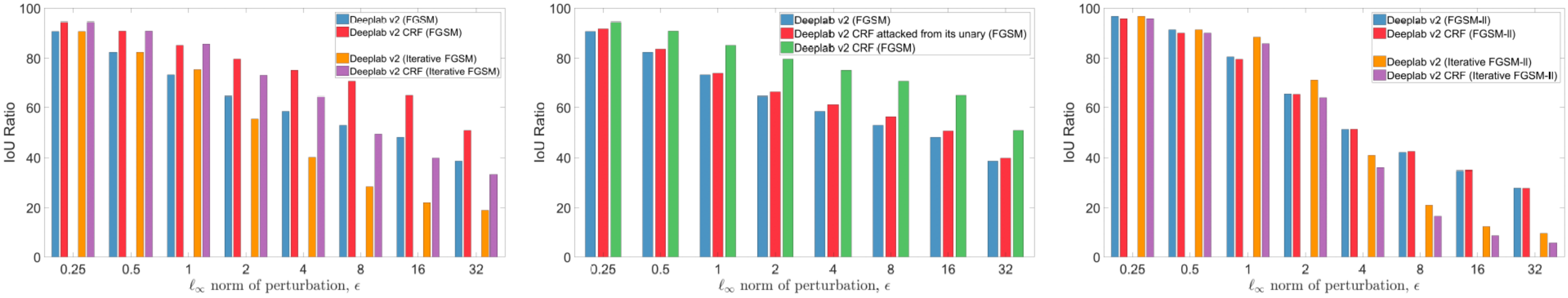
- Effect of CRFs on Adversarial Robustness



(a) (b) (c)

Figure 9: (a) On untargetted attacks on Pascal VOC, CRF-RNN is noticably more robust than FCN8s. (b) CRF-RNN is more vulnerable to black-box attacks from FCN8, due to its "gradient masking" effect which results in ineffective white-box attacks. (c) However, the CRF does not "mask" the gradient for targeted attacks and it is no more robust than FCN8s.

Findings



(a) (b) (c)

Figure 10: Similar trends are observed for Deeplab v2, which uses the DenseCRF model as post-processing, as CRF-RNN (Fig. 9) which integrates the CRF as part of the deep network. (a) On untargetted attacks, Deeplab v2 with a CRF is noticably more robust than just the Deeplab v2 network. (b) Attacks created from the base Deeplab v2 network using FGSM are more effective than those created from Deeplab v2 with CRF. This is due to the "gradient masking" effect of mean-field inference of CRFs. (c) However, the CRF does not "mask" the gradient for targeted attacks. As a result, Deeplab v2 with a CRF is no more robust than just the Deeplab v2 network.

Findings

- Image transformations and adversarial examples

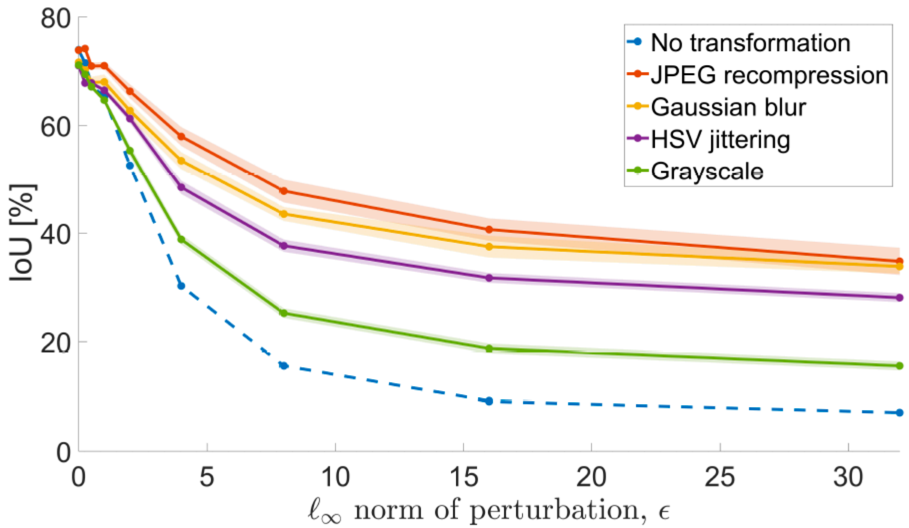


Figure 6: The adversarial examples originally generated by Iterative FGSM ll on Deeplab v2, are less malignant when the adversarial image is first pre-processed with a randomised transformation. The shaded regions correspond to two standard deviations computed from nine random trials of the randomised transformation.

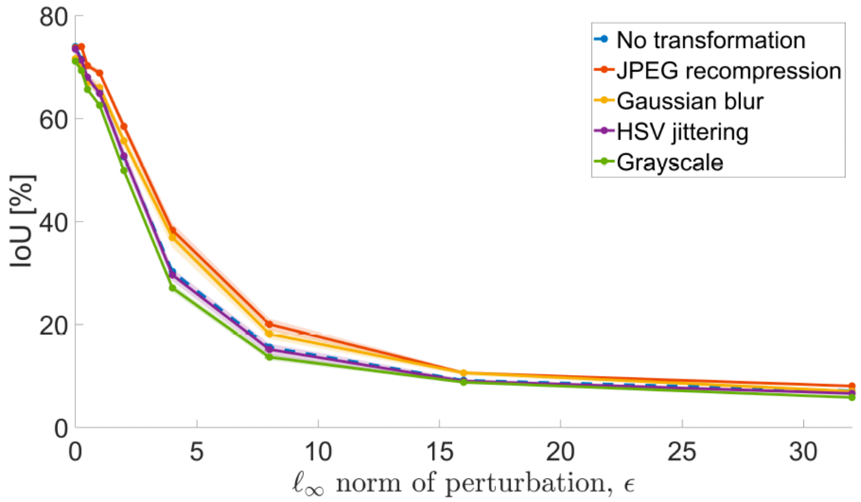


Figure 7: The randomised input transformations no longer increase the robustness of the network when the expected gradient over the distribution of the transformation functions is used in the Iterative FGSM ll attack. The shaded regions correspond to two standard deviations computed from nine random trials of the randomised transformation. The dashed blue line shows the original Iterative FGSM ll attack on non-transformed images.

Conclusion

- The paper have presented what to our knowledge is **the first rigorous evaluation of the robustness of semantic segmentation** models to adversarial attacks.
- The main observations will facilitate future efforts to understand and defend against these attacks **without compromising accuracy**.
- In the shorter term, our observations suggest that networks such as **Deeplab v2**, which is based on ResNet and performs multiscale processing, should be preferred in safety-critical applications due to their **inherent robustness**.
- They have made **numerous observations and raised questions** that will aid future work in understanding adversarial examples and developing more effective defenses.