

On the Robustness of Semantic Segmentation Models to Adversarial Attacks

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

Abstract—Deep Neural Networks (DNNs) have demonstrated exceptional performance on most recognition tasks such as image classification and segmentation. However, they have also been shown to be vulnerable to *adversarial examples*. This phenomenon has recently attracted a lot of attention but it has not been extensively studied on multiple, large-scale datasets and structured prediction tasks such as semantic segmentation which often require more specialised networks with additional components such as CRFs, dilated convolutions, skip-connections and multiscale processing.

In this paper, we present what to our knowledge is the first rigorous evaluation of adversarial attacks on modern semantic segmentation models, using two large-scale datasets. We analyse the effect of different network architectures, model capacity and multiscale processing, and show that many observations made on the task of classification do not always transfer to this more complex task. Furthermore, we show how mean-field inference in deep structured models, multiscale processing (and more generally, input transformations) naturally implement recently proposed adversarial defenses. Our observations will aid future efforts in understanding and defending against adversarial examples. Moreover, in the shorter term, we show how to effectively benchmark robustness and show which segmentation models should currently be preferred in safety-critical applications due to their inherent robustness.

Index Terms—adversarial attacks, semantic segmentation, deep learning, convolutional neural networks, machine learning security



1 INTRODUCTION

Computer vision has progressed to the point where Deep Neural Network (DNN) models for most recognition tasks such as classification or segmentation have become a widely available commodity. State-of-the-art performance on various datasets has increased at an unprecedented pace, and as a result, these models are now being deployed in more and more complex systems. However, despite DNNs performing exceptionally well in absolute performance scores, they have also been shown to be vulnerable to *adversarial examples* – images which are classified incorrectly (often with high confidence), although there is only a minimal perceptual difference with correctly classified inputs [10], [28], [81].

This raises doubts about DNNs being used in safety-critical applications such as driverless vehicles [46] or medical diagnosis [30] since the networks could inexplicably classify a natural input incorrectly although it is almost identical to examples it has classified correctly before (Fig. 1). Moreover, it allows for the possibility of malicious agents attacking systems that use neural networks [32], [52], [69], [76]. Hence, the robustness of networks perturbed by adversarial noise may be as important as the predictive accuracy on clean inputs. And if multiple models achieve comparable performance, we should always consider deploying the one which is inherently most robust to adversarial examples in (safety-critical) production settings.

This phenomenon has recently attracted a lot of attention and numerous strategies have been proposed to train DNNs to be more robust to adversarial examples [38], [53], [61], [72]. However, these defenses are not universal; they have

frequently been found to be vulnerable to other types of attacks [16], [17], [18], [44] and/or come at the cost of performance penalties on clean inputs [19], [40], [61]. To the best of our knowledge, adversarial examples have not been extensively analysed beyond standard image classification models, and often on small datasets such as MNIST or CIFAR-10 [40], [61], [72]. Hence, the vulnerability of modern DNNs to adversarial attacks on more complex tasks such as semantic segmentation in the context of real-world datasets covering different domains remains unclear.

In this paper, we present what to our knowledge is the first rigorous evaluation of the robustness of semantic segmentation models to adversarial attacks. We focus on semantic segmentation, since it is a significantly more complex task than image classification [8]. This has also been witnessed by the fact that state-of-the-art semantic segmentation models are typically based on standard image classification architectures [43], [51], [77], extended by additional components such as dilated convolutions [21], [88], specialised pooling [22], [90], skip-connections [58], Conditional Random Fields (CRFs) [1], [91] and/or multiscale processing [20], [22] whose impact on the robustness has never been thoroughly studied.

A preliminary version of our paper was presented earlier in [2], where we showed a number of interesting observations. First, we analysed the robustness of various DNN architectures to adversarial examples and showed that the Deeplab v2 network [22] was significantly more robust than approaches which achieved better prediction scores on public benchmarks [90]. Second, we showed that adversarial examples were less effective when processed at different scales. Furthermore, multiscale networks were more robust to multiple different attacks and white-box attacks on them produced more transferable perturbations.

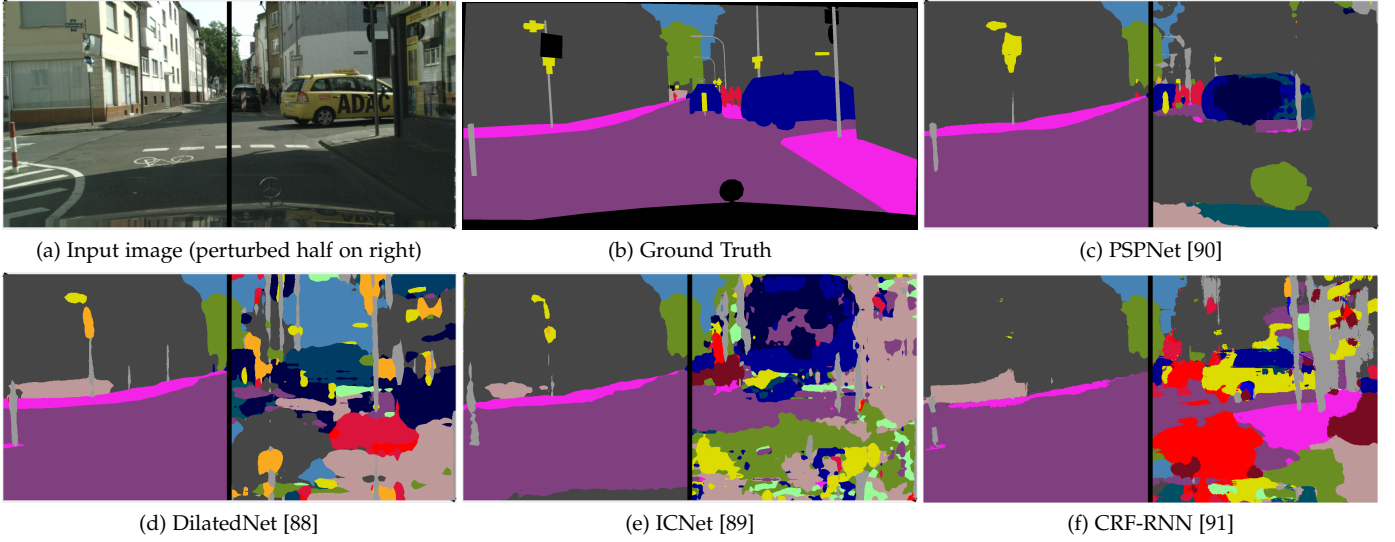


Fig. 1: The left hand side shows the original image, and the right the output when modified with imperceptible adversarial perturbations. There is a large variance in how each network’s performance is degraded, even though the perturbations are created individually for each network with the same ℓ_∞ norm of 4. We rigorously analyse a diverse range of state-of-the-art segmentation networks, observing how architectural properties, such as residual connections, multiscale processing and CRFs, and input transformations, all influence adversarial robustness. These observations will help future efforts to understand and defend against adversarial examples, whilst in the short term they suggest which networks should currently be preferred in safety-critical applications.

Third, we showed that structured prediction models had a similar effect as “gradient-masking” defense strategies [70], [72]. As such, mean field CRF inference increases robustness to untargeted adversarial attacks, but in contrast to the gradient masking defense, it also improves the network’s predictive accuracy. Finally, in contrast to the prior art [53], [57], our experiments were carried out on two large-scale, real-world datasets and (most of) our observations remained consistent across them.

We extend the initial version of our paper substantially. Inspired by the effect of multiscale processing, we examine other input transformations which neural networks are not invariant to and show that they are markedly more robust to transformed adversarial examples. However, we also show that this is true only when the attack generation process does not take knowledge of these input transformations into account; otherwise, the robustness improvements are rather marginal. These observations have important implications on producing effective physical adversarial examples in the real world. Moreover, we also show that proposed adversarial defenses should be evaluated prudently by using knowledge of the defense mechanism in the white-box attack to test it, which was not previously done in [24], [41], [54], [84]. Finally, we have updated our initial findings by showing how they are corroborated by subsequent or concurrent work to our original conference paper.

We believe our findings will facilitate future efforts in understanding and defending against adversarial examples without compromising predictive accuracy.

2 ADVERSARIAL EXAMPLES

Adversarial perturbations cause a classifier to change its original prediction, when added to the original input \mathbf{x} . For a classifier f parametrised by θ that maps $\mathbf{x} \in \mathbb{R}^m$ to y , a

target class from $\mathcal{C} = \{1, 2, \dots, C\}$, a targeted adversarial attack causes the classifier to predict y_t instead, where y_t is chosen by the attacker and $y_t \neq y$. An untargeted adversarial attack causes the classifier to predict any label besides the original prediction (from the label set $\mathcal{C} \setminus \{y\}$).

This phenomenon was initially studied in the context of malware detection and spam classification [10], [28], and has recently become popular in the context of computer vision. Szegedy *et al.* [81] defined an adversarial perturbation \mathbf{r} as the solution to the optimisation problem defining a targeted attack

$$\arg \min_{\mathbf{r}} \|\mathbf{r}\|_2 \quad \text{subject to} \quad f(\mathbf{x} + \mathbf{r}; \theta) = y_t, \quad (1)$$

where y_t is the target label of the adversarial example $\mathbf{x}^{adv} = \mathbf{x} + \mathbf{r}$. For clarity of exposition, we consider only a single label y . This naturally generalises to the case of semantic segmentation where networks are trained with an independent cross-entropy loss at each pixel.

Constraining the neural network to output y is difficult to optimise. Hence, [81] added an additional term to the objective based on the loss function used to train the network

$$\arg \min_{\mathbf{r}} \lambda \|\mathbf{r}\|_2 + L(f(\mathbf{x} + \mathbf{r}; \theta), y_t). \quad (2)$$

Here, L is the loss function between the network prediction and desired target, and λ is a positive scalar. Szegedy *et al.* [81] solved this using L-BFGS, and [18] and [23] have proposed further advances using surrogate loss functions. However, this method is computationally very expensive as it requires several minutes to produce a single attack. Hence, the following methods are used in practice:

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

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

This is a single-step, untargeted attack, which approximately minimises the ℓ_{∞} norm of the perturbation bounded by the parameter ϵ .

FGSM II [53]. This single-step attack encourages the network to classify the adversarial example as y_t by assigning

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

We follow the convention of choosing the target class as the least likely class predicted by the network [53]. We consider other choices of the target class in the supplementary material.

Iterative FGSM [53], [61]. 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

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

The $\text{clip}(\mathbf{a}, \epsilon)$ function makes sure that each element a_i of \mathbf{a} is in the range $[a_i - \epsilon, a_i + \epsilon]$. This ensures that the max-norm constraint of each component of the perturbation \mathbf{r} , being no greater than ϵ is maintained. It thus corresponds to projected gradient descent [61], with step-size α , into an ℓ_{∞} ball of radius ϵ around the input \mathbf{x} .

Iterative FGSM II [53]. This is a stronger version of FGSM II. This attack sets the target to be the least likely class predicted by the network, y_{ll} , in each iteration

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

The aforementioned attacks were all proposed in the context of image classification, but they have been adapted to the problems of semantic segmentation [23], [35], object detection [85] and visual question answering [87]. Similar, gradient-based attacks have also been proposed to minimise the ℓ_2 norm of the adversarial perturbation, \mathbf{r} , [18], [66], and also to attack other classification algorithms such as SVMs [10]. Methods to optimise the non-differentiable ℓ_0 norm of the perturbation have also been proposed [68], [71], [79].

3 ADVERSARIAL DEFENSES AND EVALUATIONS

Liu *et al.* [57] have thoroughly evaluated the transferability of adversarial examples generated on one network and tested on another unknown model, *i.e.* only as “black-box” attacks [63], [65], [70], [81]. Kurakin *et al.* [53], contrastingly, studied the adversarial training defense, which generates adversarial examples online and adds them into the training set [38], [61], [82]. They found that training with adversarial examples generated by single-step methods conferred robustness to other single-step attacks with negligible performance difference to normally trained networks on clean inputs. However, the adversarially trained network was still

as vulnerable to iterative attacks as standard models. Madry *et al.* [61], conversely, found robustness to iterative attacks by adversarial training with them. However, this was only on the small MNIST dataset. The defense was not effective on CIFAR-10, underlining the importance of testing on multiple datasets. Tramer *et al.* [82] also found that adversarially trained models were still susceptible to black-box, single-step attacks generated from other networks. Other adversarial defenses based on detecting the perturbation in the input [34], [39], [62], [78], [86] or pre-processing the input [41], [54], [75], [84] have also all been subverted [4], [5], [16], [17], [44], [83]. Recently, progress has been made on formal verification of neural networks [14], [47] which can provably compute the adversarial perturbation with the minimum norm for a network. However, as these methods are limited to certain architectures, and do not scale to large networks, they cannot be used on the state-of-the-art networks we consider in this work.

Currently, no effective defense to all adversarial attacks exist. This motivates us, for the first time to our knowledge, to study the properties of state-of-the-art segmentation networks and how they affect robustness to various adversarial attacks. Previous evaluations have only considered standard classification networks (Inception in [53], and GoogleNet, VGG and ResNet in [57]). We consider the more complex task of semantic segmentation, and evaluate eight different architectures, some of them with multiple classification backbones, and show that some features of semantic segmentation models (such as CRFs and multi-scale processing) naturally implement recently proposed adversarial defenses. Moreover, our evaluation is carried out on two large-scale datasets instead of only ImageNet as [53], [57]. This allows us to show that not all previously observed empirical results on classification transfer to segmentation.

The conclusions from our evaluations may thus aid future efforts to develop defenses to adversarial attacks that preserve predictive accuracy. Moreover, our results suggests which state-of-the-art models for semantic segmentation should currently be preferred in (safety-critical) settings where both accuracy and robustness are a priority.

Note that adversarial examples have been shown to exist for semantic segmentation before by [23], [64], [85]. However, our work is complementary, as we thoroughly study the properties of semantic segmentation networks and how they affect robustness to adversarial attacks. Previous works were not as systematic as they only considered one particular network, did not limit the norm of the adversarial perturbation and did not show how different architectural components impact adversarial robustness. Moreover, although [85] propose a gradient-based attack algorithm which considers each pixel independently, we show that similar and more common FGSM-based methods [38], [53], [61] (which [85] did not use as a baseline) are still effective.

4 EXPERIMENTAL SET-UP

We describe the datasets, DNN models, adversarial attacks and evaluation metrics used for our evaluation in this section. Exhaustive details are included in the supplementary. We have also released our code¹ to aid reproducibility.

1. www.robots.ox.ac.uk/~aarnab/adversarial_robustness.html

Datasets.

We use the Pascal VOC [31] and Cityscapes [25] validation sets, the two most widely used semantic segmentation benchmarks. Pascal VOC consists of internet-images labelled with 21 different classes. The reduced validation [58], [91] set contains 346 images, and the training set has about 70000 images when combined with additional annotations from [42] and [56]. Cityscapes consists of road-scenes captured from car-mounted cameras and has 19 classes. The validation set has 500 images, and the training set totals about 23000 images. As this dataset provides high-resolution imagery (2048×1024 pixels) which require too much memory for some models, we have resized all images to 1024×512 when evaluating.

Models.

We use a wide variety of current or previous state-of-the-art models, ranging from lightweight networks suitable for embedded applications to complex models which explicitly enforce structural constraints. Whenever possible, we have used publicly available code or trained models. The models we had to retrain achieve similar performance to the ones trained by the original authors.

We used the public models of CRF-RNN [91], DilatedNet [88], PSPNet [90] on Cityscapes, ICNet [90] and SegNet [7]. We retrained FCN [58] and E-Net [73], as well as Deeplab v2 [22] and PSPNet for VOC as the public models are trained with the validation set. Our selection of networks are based on both VGG [77] and ResNet [43] backbones, whilst E-Net and ICNet employ custom architectures for real-time applications whose parameters measure only 1.5MB and 30.1MB in 32-bit floats, respectively. Furthermore, the models we evaluate use a variety of unique approaches including dilated convolutions [22], [88], skip-connections [58], specialised pooling [22], [90], encoder-decoder architecture [7], [73], multiscale processing [22] and CRFs [91]. In all our experiments, we evaluate the model in the same manner it was trained – CRF post-processing or multiscale ensembling is not performed unless the network incorporated CRFs [91] or multiscale averaging [22] as network layers whilst training.

Adversarial attacks.

We use the FGSM, FGSM II, Iterative FGSM and Iterative FGSM II attacks described in Sec. 2. Kurakin *et al.* [53] set the number of iterations of iterative attacks to $\min(\epsilon + 4, \lceil 1.25\epsilon \rceil)$. However, we found that attacks did not always converge with this setting, and instead used $\max(\epsilon + 4, \lceil 5\epsilon \rceil)$. We set our step-size $\alpha = \min(1, \epsilon)$ meaning that the value of each pixel is changed by α (if it is not clipped due to the max-norm constraint) every iteration. The Iterative FGSM (untargeted) and FGSM II (targeted) attacks are only reported in the supplementary as we observed similar trends on FGSM and Iterative FGSM II. We evaluated these attacks when setting the ℓ_∞ norm of the perturbations ϵ to each value from $\{0.25, 0.5, 1, 2, 4, 8, 16, 32\}$. Even small values such as $\epsilon = 0.25$ introduce errors among all the models we evaluated. The maximum value of ϵ was chosen as 32 since the perturbation was conspicuous at this point. Qualitative examples of these attacks are shown in the supplementary.

Evaluation metric.

The Intersection over Union (IoU) is the primary metric used in evaluating semantic segmentation [25], [31]. However, as the accuracy of each model varies, we adapt the relative metric used by [53] for image classification and measure adversarial robustness using the *IoU Ratio* – the ratio of the network’s IoU on adversarial examples to that on clean images computed over the entire dataset. As the relative ranking between models for the IoU Ratio and absolute IoU is typically the same, we report the latter only in the supplementary.

5 THE ROBUSTNESS OF DIFFERENT ARCHITECTURES

In this section, we evaluate the robustness of different network architectures. Our experiments show a more nuanced relationship between model capacity and adversarial robustness, by considering a different setting to the previous findings of [53], [61]. Additionally, our results also support why JPEG compression as a pre-processing step mitigates small perturbations [29].

5.1 The robustness of different networks

Fig. 2 shows the robustness of several state-of-the-art models on the VOC dataset. In general, ResNet-based models not only achieve higher accuracy on clean inputs but are also more robust to adversarial inputs. This is particularly the case for the single-step FGSM attack (Fig. 2a). On the more effective Iterative FGSM II attack, the margin between the most and least robust network is smaller as none of them perform well (Fig. 2b). However, we note that iterative attacks tend not to transfer to other models [53] (Sec. 6.2). Thus, they are less useful in practical, black-box attacks.

In particular, we have evaluated the FCN8s [58] and Deeplab-v2 with ASPP [22] models based on the popular VGG-16 [77] and ResNet-101 [43] backbones. In both cases, the ResNet variant shows greater robustness. We also observe that most of the networks achieve similar scores on clean inputs. As a result, the relative rankings of models in Fig. 2 for the IoU Ratio is about the same as their ranking on clean inputs. Furthermore, the best performing model on clean inputs, PSPNet [90] is actually less robust than Deeplab v2 with Multiscale ASPP [22]: For all ϵ values we tested, the absolute IoU score of Deeplab v2 was higher than PSPNet. These observations as well as results on FGSM II and Iterative FGSM showing that the relative ranking of robustness for the different networks is similar, are detailed in the supplementary material.

5.2 Model capacity and residual connections

Madry *et al.* [61] and Kurakin *et al.* [53] have studied the effect of model capacity on adversarial robustness by changing the number of filters at each convolutional layer in their network, since they used the parameter count as a proxy for the model capacity. Madry *et al.* [61] observed, on MNIST and CIFAR-10, that networks trained on clean examples with a small number of parameters are the most vulnerable to adversarial examples. This observation, which suggests

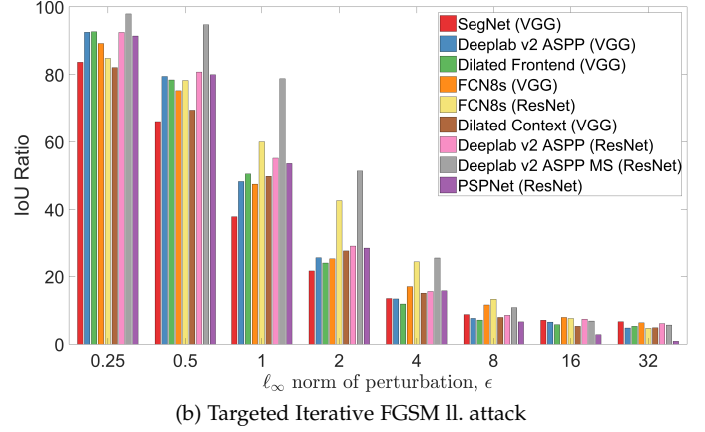
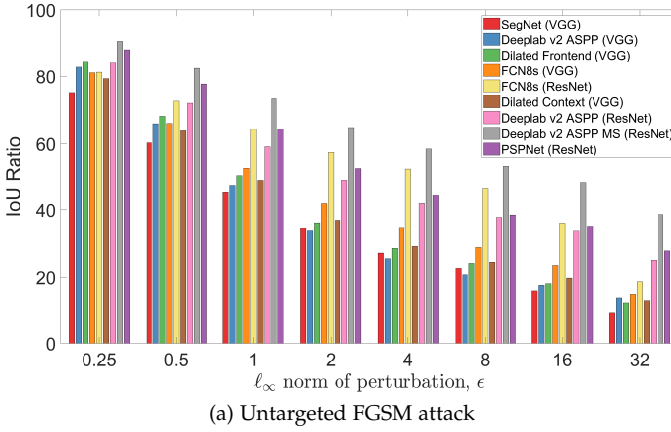


Fig. 2: Adversarial robustness of state-of-the-art models on Pascal VOC. Models based on the ResNet backbone tend to be more robust. For instance, FCN8s and Deeplab v2 ASPP with a ResNet-101 backbone are more robust than with the VGG backbone. Moreover, as expected, the Iterative FGSM II attack is more powerful at fooling networks than single-step FGSM. Models are ordered by increasing IoU on clean inputs. Results on additional attacks are in the supplementary.

that small networks with few parameters are the most vulnerable to adversarial examples, would have serious safety implications on the deployment of lightweight models, typically required in robotics, autonomous vehicles and embedded system applications. Here, we instead analyse different network structures that are used in practice (unlike [61] and [53] who used the same architecture with a different number of filters) and show in Fig. 3 that lightweight networks such as E-Net [73] (only 1.5 MB) and IC-Net [89] (only 30.1 MB) are affected by adversarial examples similarly as Dilated-Net [88] which has 512.6 MB in parameters (using 32-bit floats). Dilated-Net is only more robust than both of these lightweight networks for FGSM and FGSM-II with $\epsilon \geq 4$ (which is also when perturbations become visible to the naked eye). Note that both E-Net and IC-Net have custom backbones and heavily use residual connections.

Fig. 3 also shows that adding the “Context” module of Dilated-Net onto the “Front-end” slightly reduces robustness across all ϵ values on both attacks on Cityscapes. Fig. 2 shows that this is observed for most ϵ values on VOC as well. This is even though the additional parameters of the “Context” module increases accuracy on clean inputs. Whilst models with higher capacity may be more resistant to adversarial attacks (as posited by Madry *et al.* [61]), one cannot compare the capacities of different networks, given that neither the most accurate network (PSPNet) or the network with the most parameters (Dilated-Net) are actually the most robust.

5.3 The unexpected effectiveness of single-step methods on Cityscapes

The single-step FGSM and FGSM II attacks are significantly more effective on Cityscapes than on Pascal VOC. The IoU ratio for FGSM at $\epsilon = 32$ for PSPNet, Dilated Context and FCN8s is 2.5%, 2.8% and 8.0%, respectively, on Cityscapes. On Pascal VOC, it is substantially higher at 27.9%, 12.2% and 15.0%. As expected, the iterative methods still significantly outperform single-step methods across both datasets.

Thus, it may be a dataset property that causes the network to learn weights more susceptible to single-step

attacks. Cityscapes has, subjectively, less variability than VOC and it also labels “stuff” classes [36]. The effect of the training set on adversarial attacks has not been considered before, and most prior work used MNIST [38], [61], [81] or ImageNet [53], [57], [82]. However, [11] and [49], showed that the test error of an SVM and neural network could respectively be increased by inserting “poisonous” examples into its training set. Results from the FGSM II attack, which shows the same trend as FGSM, are in the supplementary.

5.4 Imperceptible perturbations

With $\epsilon = 0.25$, the perturbation is so small that the RGB values of the image pixels (assuming integers $\in [0, 255]$) are usually unchanged. Nevertheless, Fig. 2 and 3 show that the performance of all analysed models were degraded by at least 3% relative IoU for each attack. The observation of [29], that lossy JPEG as a pre-processing step helps to mitigate FGSM for small ϵ is thus not surprising as JPEG does not entirely preserve these small, high-frequency perturbations and the result is also finally rounded to integers.

5.5 Relation with concurrent work

Our results are also corroborated by the concurrent work of Cubuk *et al.* [26] who performed Neural Architecture Search to find architectures that are more robust to adversarial examples. Cubuk *et al.* [26] found that their best architecture had more identity connections and depth than their baseline. This agrees with our observation that models based on ResNet typically have higher robustness and accuracy on clean inputs.

The authors also observed a correlation between accuracy on clean data and robustness. We also observed this correlation (Fig. 4), although the most accurate model on clean inputs (PSPNet) is not the most robust (Deeplab v2 Multiscale). Figure 4 shows the results on the FGSM attack at $\epsilon = 8$, for consistency with [26].

5.6 Discussion

We have shown that models with residual connections (ResNet, E-Net, ICNet) are inherently more robust than

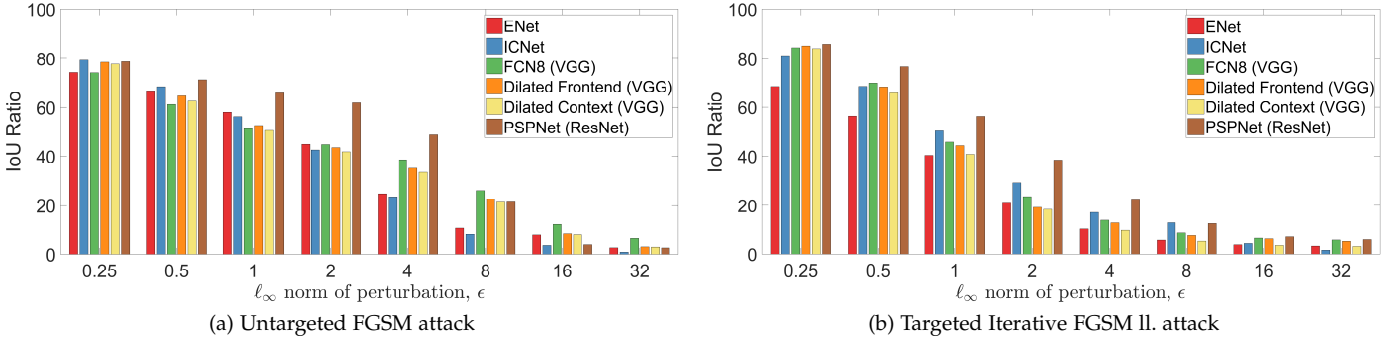


Fig. 3: Adversarial robustness of state-of-the-art models on the Cityscapes dataset. We observe that lightweight networks such as E-Net [73] and ICNet [89] are often about as robust as Dilated-Net [88] ($341\times$ more parameters than E-Net). Dilated-Net without its “Context” module is also slightly more robust than the full network (these findings regarding parameter count are contrary to Madry *et al.* [61] who however did not evaluate different architectures). Both attacks are very effective after $\epsilon \geq 16$, with performance of all networks degraded considerably. As with the VOC dataset, ResNet (PSPNet) architectures are more robust than VGG (Dilated-Net and FCN8).

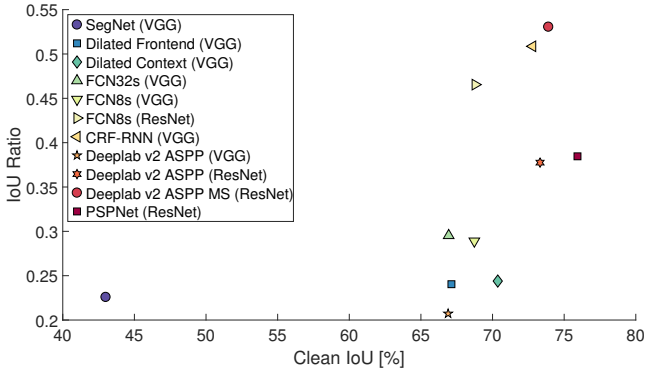


Fig. 4: The IoU Ratio compared to the IoU on clean inputs on the Pascal VOC dataset, for the FGSM attack with $\epsilon = 8$. The relative ordering of the models is the same if we plot the absolute IoU on adversarial inputs, with the exception of SegNet which is then ranked the lowest.

chain-like VGG-based networks, even if the number of parameters of the VGG model is orders of magnitude larger. Moreover, Dilated-Net, without its “Context” module is more robust than its more performant, full version. This is contrary to the observations regarding parameter count of [61], who noted that smaller networks were less adversarially robust on MNIST and CIFAR-10. However, a key difference between our experiments and [53], [61] is that we have considered different network architectures whilst [61] only changed the number of filters at each DNN layer. Our results in this regard are more in line with Kurakin *et al.* [53] who reported with Inception-v3 [80] based architectures on ImageNet that models that were too large or too small were less adversarially robust. The most robust model was Deeplab v2 with Multiscale ASPP, outperforming the current state-of-the-art PSPNet [90], in absolute IoU on adversarial inputs.

We also found that perturbations that do not even change the image’s integral RGB values still degraded performance of all models, and that single-step attacks are significantly more effective on Cityscapes than VOC, achieving as low as 0.8% relative IoU, raising questions about how the

training data of a network affects its decision boundaries. Also, explaining the effect of residual connections on adversarial robustness remains an open research question. As Deeplab v2 showed a significant increase in robustness over its single-scale variant, we analyse the effects of multiscale processing next in Sec. 6. Thereafter, we study CRFs, a common component in semantic segmentation models.

6 MULTISCALE PROCESSING AND TRANSFERABILITY OF ADVERSARIAL EXAMPLES

Deeplab v2 with Multiscale ASPP was the most robust model to various attacks in Sec. 5, with a significant difference to its single-scale variant. In this section, we first examine the effect of multiscale processing and then relate our observations to concurrent work.

6.1 Multiscale processing

The Deeplab v2 network processes images at three different resolutions, 50%, 75% and 100% where the weights are shared among each of the scale branches. The results from each scale are upsampled to a common resolution, and then max-pooled such that the most confident prediction at each pixel from each of the scale branches is chosen [22]. This network is trained in this multiscale manner, although it is possible to perform this multiscale ensembling as a post-processing step at test-time only [21], [27], [55], [90].

We hypothesise that adversarial attacks, when generated at a single scale, are no longer as malignant when processed at another. This is because CNNs are not invariant to scale, and a range of other transformations [33], [45], [74]. And although it is possible to generate adversarial attacks from multiple different scales of the input, these examples may not be as effective at a single scale, making networks which process images at multiple scales more robust. We investigate the transferability of adversarial perturbations generated at one scale and evaluated at another in Sec. 6.2, and the robustness and transferability of multiscale networks in Sec. 6.3. Thereafter, we relate our findings to concurrent work.

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 II ($\epsilon = 8$) | | | |
|--------------------------------|-------------------------|-------------|-------------|-------------|--------------------------------------|-------------|-------------|-------------|
| | 50% | 75% | 100% | Multiscale | 50% | 75% | 100% | Multiscale |
| Deeplab v2 50% scale (ResNet) | <u>37.3</u> | 70.5 | 84.8 | 60.3 | 18.0 | 92.0 | 96.9 | 20.0 |
| Deeplab v2 75% scale (ResNet) | 85.5 | <u>39.7</u> | 62.2 | 50.8 | 99.5 | <u>17.9</u> | 89.9 | 20.4 |
| Deeplab v2 100% scale (ResNet) | 93.6 | 57.9 | <u>37.7</u> | 37.2 | 100.0 | 79.0 | <u>15.5</u> | 16.8 |
| Deeplab v2 Multiscale (ResNet) | 83.7 | 57.6 | 62.3 | <u>53.1</u> | 99.6 | 90.2 | <u>91.9</u> | <u>21.5</u> |
| 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 |

6.2 The transferability of adversarial examples at different scales

Table 1 shows results for the FGSM and Iterative FGSM II attacks. The diagonals show “white-box” attacks where the adversarial examples are generated from the attacked network. These attacks typically result in the greatest performance degradation, as expected. The off-diagonals show the transferability of perturbations generated from other networks. In contrast to Iterative FGSM II, FGSM attacks transfer well to other networks, which confirms the observations [53] made in the context of image classification.

The attack produced from 50% resolution inputs transfers poorly to other scales of Deeplab v2 and other architectures, and vice versa. This is seen by looking across the columns and rows of Tab. 1 respectively. All other models, FCN (VGG and ResNet) and Deeplab v2 VGG were trained at 100% resolution, and Tab. 1 shows that perturbations generated from the multiscale and 100% resolutions of Deeplab v2 transfer the best. This supports the hypothesis that adversarial attacks produced at one scale are not as effective when evaluated at another since CNNs are not scale invariant (the network activations change considerably).

6.3 Multiscale networks and adversarial examples

The multiscale version of Deeplab v2 is the most robust to white-box attacks (Tab. 1, Fig. 2) as well as perturbations generated from single-scale networks. Moreover, attacks produced from it transfer the best to other networks as well, as shown by the bolded entries. This is probably because attacks generated from this model are produced from multiple input resolutions simultaneously. For the Iterative FGSM II attack, only the perturbations from the multiscale version of Deeplab v2 transfer well to other networks, achieving a similar IoU ratio as a white-box attack. However, this is only the case when attacking a different scale of Deeplab. Whilst perturbations from multiscale Deeplab v2 transfer better on FCN than from single-scale inputs, they are still far from the efficacy of a white-box attack (which has an IoU ratio of 15.2% on FCN-VGG and 26.4% on FCN-ResNet).

Adversarial perturbations generated from multiscale inputs to FCN8 (which has only been trained at a single scale) behave in a similar way: FCN8 with multiscale inputs is more robust to white-box attacks, and its perturbations transfer better to other networks. This suggests that the observations seen in Tab. 1 are not properties of training

the network, but rather the fact that CNNs are not scale invariant. Furthermore, an alternative to max-pooling the predictions at each scale is to average them. Average-pooling produces similar results to max-pooling. Details of these experiments, along with results using different attacks and l_∞ norms (ϵ values), are presented in the supplementary.

6.4 Relation to other defenses

Our observations relate to the “random resizing” defense of [84] in concurrent work. Here, the input image is randomly resized and then classified. This defense exploits (but does not attribute its efficacy to) the fact that CNNs are not scale invariant and that adversarial examples were only generated at the original scale. Our findings suggest that this defense (which is very similar to the multiscale processing performed naturally by Deeplab v2) could be defeated by creating adversarial attacks from multiple scales, as done in this work, and this has indeed been verified [5], [83].

7 IMAGE TRANSFORMATIONS AND ADVERSARIAL EXAMPLES

In Sec. 6, we posited that adversarial examples are less malicious when processed at different scales since CNNs are not scale invariant. Scale changes are used in segmentation architectures to recognise objects at different resolutions, however, this is not the only commonly used image transformation. In this section, we consider a number of other common input transformations, and examine their effect on adversarial robustness of CNNs for semantic segmentation.

In the following, each transformation is applied to the input image before it is processed by the neural network and we examine how it affects the robustness to adversarial examples. Following on from Sec. 6, we use the Deeplab v2 MS network, which we found to be the most robust in Sec. 5, and consider the following four transformations (illustrated in Fig. 5) which are ubiquitous in computer vision and image processing:

JPEG recompression. The image is compressed using JPEG with a “quality” parameter drawn randomly between 50% and 100%. The image is then reconstructed and processed by the network.

Gaussian blur. The input image is blurred by a Gaussian filter with a bandwidth uniformly drawn from $[0, 2]$, which ensures that all objects in the image are still recognisable and can be segmented precisely.

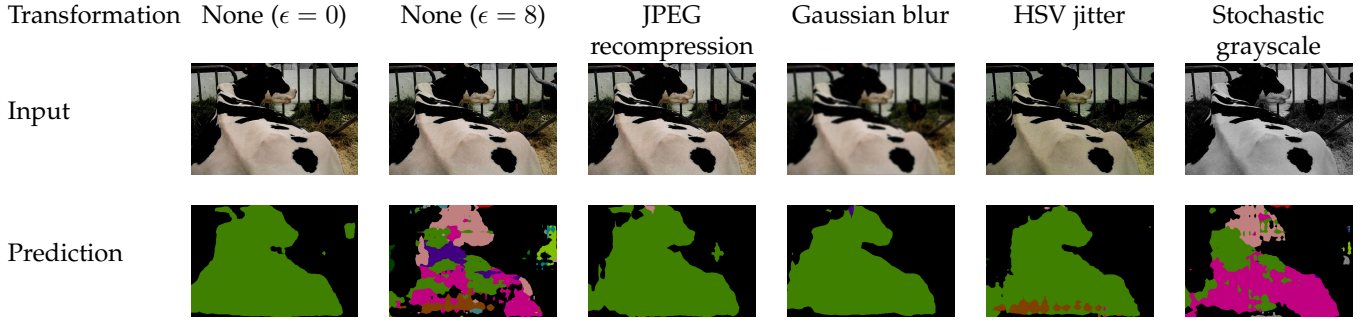


Fig. 5: Input transformations of adversarial examples generated by Iterative FGSM II (Eq. 6) significantly change the prediction of the Deeplab v2 network. These input transformations, however, barely change the output when they are applied to clean images. The l_∞ norm of the perturbation, $\epsilon = 8$, is visible when looking carefully on screen.

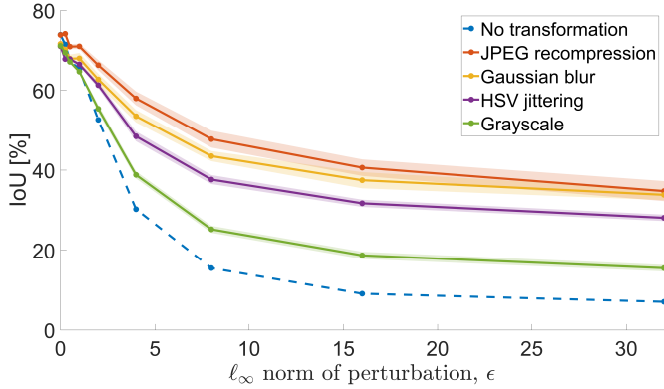


Fig. 6: The adversarial examples originally generated by Iterative FGSM II on Deeplab v2, are less malignant when the adversarial image is first pre-processed with a randomised transformation. The shaded regions correspond to two standard deviations computed from nine random trials of the randomised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

HSV jitter. The image is converted to the HSV colour space (which is more perceptually similar than the RGB space). Next, each pixel is perturbed by a value drawn uniformly between $[-30, 30]$ and then converted back to RGB space for processing.

Grayscale. The input image is converted to grayscale by setting all three image channels to have the same value at each pixel. This was performed using a convex combination of each of the three RGB channels, with each of the co-efficients sampled from a flat Dirichlet distribution.

Note that none of the transformations affect the image spatial co-ordinates, which means that it is suitable for using with semantic segmentation models without any additional post-processing. These transformations, though quite disparate, all have a similar effect on adversarial robustness as described in the next subsection.

7.1 Robustness conferred by randomised input transformations

Figure 6 shows that each type of input transformation substantially increases the robustness of Deeplab v2 to the Iterative FGSM II attack on the VOC dataset, with “JPEG recompression” and “Gaussian blur” providing substantial

benefits. Converting the image to grayscale with random channel coefficients provides a smaller, but still sizeable, improvement. These findings are consistent and show little variance over 9 different trials, since each input transformation is randomised. The IoU of the transformed images at $\epsilon = 0$ (i.e. corresponding to no attack) is similar to the original image with the largest difference about 2%. Therefore, the network is more sensitive to input transformations on adversarial images than it is on clean ones.

These results, in addition to Tab. 6, show that as neural networks are not invariant to many classes of transformations of the input, their predictions on adversarial examples subject to these transformations change. Consequently, predictions on transformed adversarial inputs are different to the original adversarial example, and this typically results in the adversarial example becoming less malignant. These findings are consistent across a broad range of geometric and photometric transformations.

Dziugaite *et al.* [29] previously observed that JPEG recompression improved adversarial robustness for small ϵ values in the context of image classification. However, the authors hypothesised that a special property of the JPEG algorithm (i.e. mapping images back onto the manifold of natural images) was the reason it conferred additional robustness. In contrast, our study of various different transformations suggest that JPEG recompression is just one instance of the numerous input transformations which neural networks are not invariant to. As a result, JPEG recompression, along with other image transformations, increases robustness to adversarial examples that were generated by attacks which did not take it into account.

7.2 Subverting randomised, non-differentiable input transformations

The results shown in Fig. 6 suggest that randomised input transformations serve as an effective defense to adversarial attacks. They significantly reduced the effectiveness of the Iterative FGSM II attack, which has been the most powerful attack in our experiments, and the result for $\epsilon = 0$ also shows that this method has minimal performance penalties on clean inputs. This reasoning has been exploited by the concurrent work of [41], where the authors showed how several different input transformations increased the robustness of image classification models to adversarial attacks.

However, the results in Fig. 6 and [41] assume that knowledge of the defence mechanism (randomised input

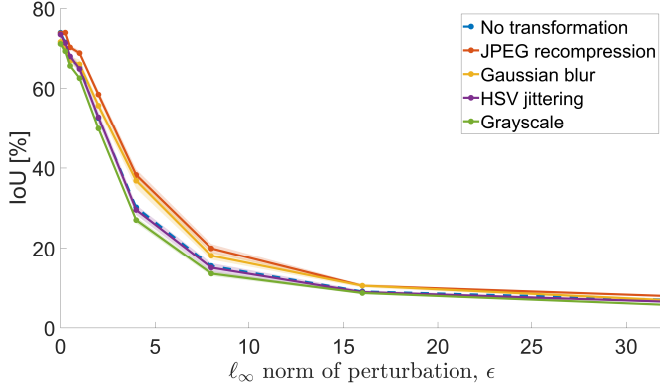


Fig. 7: The randomised input transformations no longer increase the robustness of the network when the expected gradient over the distribution of the transformation functions is used in the Iterative FGSM II attack. The shaded regions correspond to two standard deviations computed from nine random trials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

transformations in this case) is not exploited in generating the adversarial attack. This methodology goes against Kerckhoffs’ principle [48] – the basis of modern cryptographic systems – which states that a system should be secure if everything about it barring the key is public knowledge.

Consequently, to confirm if randomised input transformations really confer adversarial robustness, we modify the Iterative FGSM II update (Eq. 6) to compute the expected gradient over the distribution of transformations which could be applied at inference time,

$$\mathbf{g} = \mathbb{E}_{t \sim \mathcal{T}} [\nabla_{\mathbf{x}_t^{\text{adv}}} L(f(t(\mathbf{x}_t^{\text{adv}}); \theta), y_u)], \quad (7)$$

$$\mathbf{x}_{t+1}^{\text{adv}} = \text{clip}(\mathbf{x}_t^{\text{adv}} - \alpha \cdot \mathbf{g}, \epsilon), \quad (8)$$

where \mathcal{T} is the distribution of transformation functions t . This method uses the fact that $\nabla_{\mathbf{x}} \mathbb{E}_{t \sim \mathcal{T}} f(t(x)) = \mathbb{E}_{t \sim \mathcal{T}} \nabla_{\mathbf{x}} f(t(x))$. It has also been used by [5] to estimate the gradient of neural networks with randomised non-differentiable adversarial defences [84]. This variant of the FGSM attack corresponds to sampling from the distribution of transformations, computing the loss and gradient of the image with respect to the loss, and averaging this gradient over many samples before performing the update.

Note that some transformations, such as JPEG recompression, are not differentiable. In this case, we use the straight-through estimator [9] which assumes, when computing the gradient using backpropagation, that the transformation is the identity function.

Figure 7 shows the results of the Expectation over Transformations (EOT) attack (Eq. 8) on the Deeplab v2 model on the Pascal VOC dataset, with the expectation computed over 10 samples. The randomised JPEG and Gaussian blur input transformations increase the robustness of the model marginally, whilst jittering pixel values in the HSV space and grayscale conversion provide no additional robustness. The final IoU is similar to the original model that did not use randomised input transformations and was attacked with the standard Iterative FGSM II attack. To our knowledge, we are the first who show that neural networks can easily

be attacked with both non-differentiable and randomised input transformations. However, we point out that [5] have attacked numerous recent defenses, some which were either non-differentiable or randomised, but not both.

7.3 Transferability of input transformations

The previous two parts have shown that using input transformations reduces the malignancy of an adversarial perturbation (Sec. 7.1). Our second observation, however, showed that whenever we exploit knowledge about the input transformation during attack generation, the perturbation can become as malignant as the attack on the image with no input transformation (Sec. 7.2).

In this section, we examine the transferability of the perturbations generated from different transformations as described in Sec. 7.2. For example, we consider the efficacy of a perturbation created using the “JPEG recompression” transformation when the network’s input is pre-processed with “Gaussian blur” instead. This has important implications on the robustness and security of neural networks; if the perturbations do not transfer across different input transformations, it would suggest that a “security-through-obscurity” approach could be used, as a defender could secure their system by ensuring that the attacker does not know the input transformations they are using. It also has implications on our ability to produce malicious physical adversarial examples [52], [76], as physical objects in the real world can be viewed from a diverse range of illumination conditions, camera viewpoints and other transformations of an original canonical view.

Table 2 and Fig. 8 show our results when the adversarial perturbation generated using one distribution of transformations is applied on a network using another randomised transformation as pre-processing. Table 2 shows the absolute IoU (to account for the fact that input transformations cause slight changes on the accuracy of clean inputs) for $\epsilon = 8$, which is when the adversarial perturbations become conspicuous to the human eye, whilst Fig. 8 summarises the results for all ϵ values. Perturbations generated to target “JPEG recompression” or “Gaussian blur” input pre-processing (the two transformations which confer the most robustness to standard attacks generated without transformations (Fig. 6)), show poor transferability when the “Grayscale” or “HSV jitter” input transformation is used instead. In contrast, perturbations generated to target the “Grayscale” input transformation transfer the best to the other input transformations that we have considered in our experiments. Additionally, the last row of Tab. 2 shows that when no input transformation is used at inference time, attacks generated to target a particular input transformation are more effective with the exception of the “Grayscale” transformation. This corresponds with our results in Sec. 6 where adversarial attacks generated at multiple scales transferred better to other models.

There are clearly a myriad of input transformations that could be performed as input pre-processing to a neural network, of which we have considered only a handful. Nevertheless, it is evident that targeting some input transformations (such as grayscale conversion) appears to produce perturbations that are more transferable to other

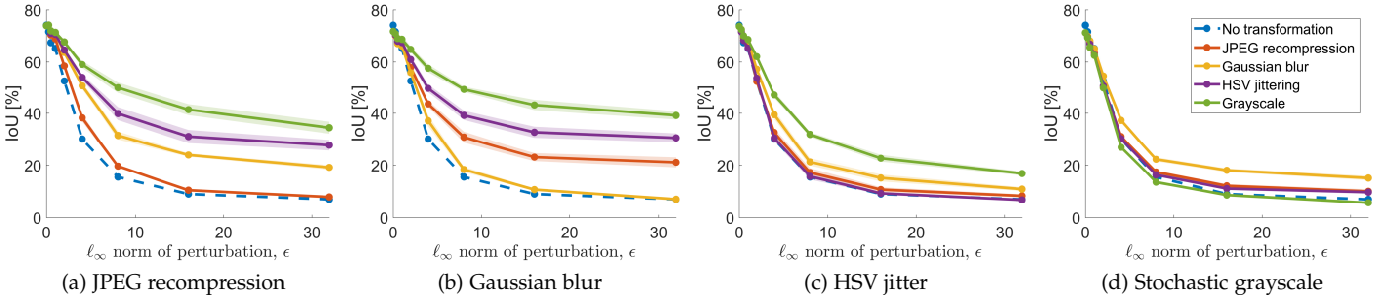


Fig. 8: The effectiveness of adversarial examples generated with one distribution of input transformations, and evaluated with another. The title of each graph shows the input transformation the adversarial examples were generated with. Each graph is effectively a column of Tab. 2 for multiple ϵ values. The dotted blue line shows the Iterative FGSM II attack when input transformations are not used at either inference or attack generation time.

TABLE 2: Transferability of adversarial attacks generated with different input transformation distributions. The left column indicates the distribution of transformations (as described in Sec. 7) that was used at inference time, and the other columns show the input transformations used when generating the attack. This table shows the mean absolute IoU scores of the Deeplab v2 network on the VOC dataset for the Iterative FGSM II attack with $\epsilon = 8$. The diagonals show “white-box” entries where the input transformation distribution used at inference time is used to generate the attack as well. The bold entries off the diagonals show the strongest attack when a different transformation distribution is used at inference time.

| Input transformation at inference time | Input transformation to generate attack | | | | |
|--|---|---------------|---------------|----------------------|-------------|
| | JPEG recompression | Gaussian blur | HSV jittering | Stochastic grayscale | None |
| JPEG recompression | 19.7 | 30.9 | 17.2 | 17.4 | 47.7 |
| Gaussian blur | 31.6 | <u>18.4</u> | 21.3 | 22.4 | 43.5 |
| HSV jitter | 39.9 | 39.2 | <u>15.7</u> | 16.3 | 33.5 |
| Stochastic grayscale | 50.0 | 49.3 | 32.0 | <u>13.6</u> | 25.2 |
| None | 11.6 | 14.4 | 12.0 | 24.4 | <u>15.5</u> |

input transformations in comparison to others (JPEG recompression). This raises an important research question about why including certain input transformations into the attack generation process transfer better to other input transformations. It also suggests another critical and open question, whether it is possible to produce adversarial perturbations that are malignant across all input transformations without modelling all of these transformations explicitly when generating the attack.

7.4 Relation to concurrent work

Our findings corroborate with concurrent work discussing producing physical adversarial attacks. Lu *et al.* [59] created adversarial traffic signs by capturing images of road signs from 0.5m and 1.5m away, generating attacks from these images on a computer, and then printing out the adversarial image onto paper. Whilst the printed image taken from 0.5m away fooled an object detector viewing the adversarial image from 0.5m, it did not when viewed from 1.5m and vice versa. This result is corroborated by Tab. 1 which shows that adversarial examples transfer poorly across different scales. Subsequent work [6], [32] has shown that it is possible to construct adversarial examples that are malignant across multiple different scales by incorporating scale changes into the attack generation process. This is again supported by our results in Tab. 1, and Sec. 7.2 which also show this effect for a number of other input transformations. When producing physical adversarial attacks, it is difficult to model all the transformations that the original image could be subject to, and as reflected by Sec. 7.3, adversarial examples generated

to target a particular transformation do not always transfer well to other input transformations. This may explain why the adversarial traffic signs generated by [32] have not been able to fool the detectors subsequently evaluated by Lu *et al.* [60]. Our observation that input transformations that were not explicitly modelled in the attack generation process mitigate the effectiveness of adversarial attacks also suggest that future work on physical adversarial attacks requires much more robust evaluation than initial work in this area [12], [32], [52], [59]. This is to ascertain whether the proposed attacks are still effective in the diverse environmental conditions that images of the adversarial object may be acquired from.

Our study of the effect of input transformations on adversarial robustness also emphasises the importance of incorporating knowledge of the proposed adversarial defence into the attack used to validate it (Kerckhoff’s principle [48]). This is not the case for many recently proposed defenses [13], [41], [54], [84] which have all subsequently been defeated [4], [5], [17], [83].

8 EFFECT OF CRFs ON ADVERSARIAL ROBUSTNESS

Conditional Random Fields (CRFs) are commonly used in semantic segmentation to enforce structural constraints [3]. The most common formulation is DenseCRF [50], which encourages nearby (in terms of position or appearance) pixels to take on the same label and hence prefers smooth labelling. This is done by a pairwise potential function,

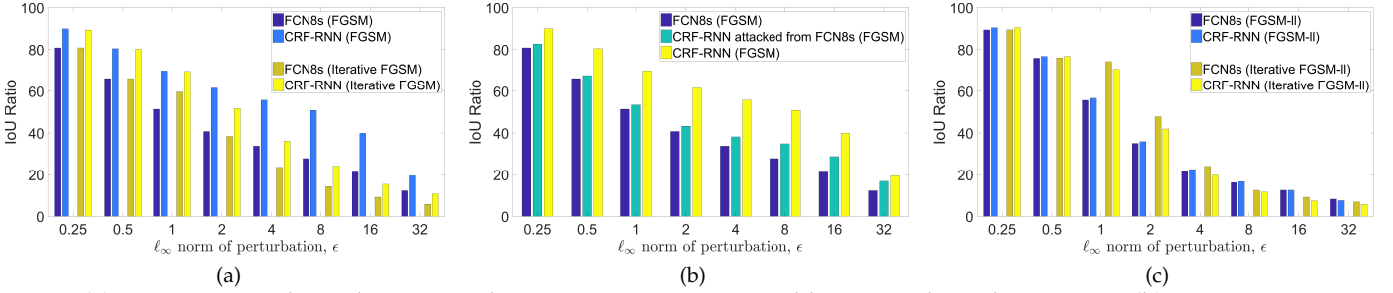


Fig. 9: (a) On untargeted attacks on Pascal VOC, CRF-RNN is noticeably more robust than FCN8s. (b) CRF-RNN is more vulnerable to black-box attacks from FCN8, due to its “gradient masking” effect which results in ineffective white-box attacks. (c) However, the CRF does not “mask” the gradient for targeted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

defined between each pair of pixels, which takes the form of a weighted sum of a bilateral and Gaussian filter.

Intuitively, one may observe that adversarial perturbations typically appear as a high-frequency noise, and thus the pairwise terms of DenseCRF which act as a low-pass filter, may provide resistance to adversarial examples. To verify this hypothesis, we consider CRF-RNN [91]. This approach formulates mean-field inference of DenseCRF as an RNN which is appended to the FCN8s network [58], enabling end-to-end training.

8.1 CRFs confer robustness to untargeted attacks

Fig. 9a shows that CRF-RNN is markedly more robust than FCN8s to the untargeted FGSM and Iterative FGSM attacks. To verify the hypothesis that the smoothing effect of the pairwise terms increases the robustness to adversarial attacks, we evaluated various values of the bandwidth hyperparameters defining the pairwise potentials (not learned; in Fig. 9a, we used the values of the public model).

Higher bandwidth values (increasing smoothness) do not actually lead to greater robustness. Instead, we observed a correlation between the final confidence of the predictions (from different hyperparameter settings) and robustness to adversarial examples. We measured confidence according to the probability of the highest-scoring label at each pixel, as well as the entropy of the marginal distribution over all labels at each pixel. The mean confidence and entropy for CRF-RNN (with original hyperparameters) is 99.1% and 0.025 nats respectively, whilst it is 95.2% and 0.13 nats for FCN8s (additional details in supplementary). The fact that mean-field inference tends to produce overconfident predictions has also been noted previously by [67] and [15].

More confident predictions lead to a smaller loss, making attacks which use the gradient of the loss with respect to the input less effective. The “Defensive Distillation” approach of [72] made use of a similar fact by increasing the confidence of the model’s predictions, resulting in gradients of smaller norm. The key difference is that CRFs increase the confidence as a by-product of a technique generally used to improve accuracy on numerous pixel-wise labelling tasks, while the effect of [72] on accuracy is unknown, as it was only tested on the saturated MNIST and CIFAR10 datasets.

8.2 Circumventing the CRF

Although CRFs are more resistant to untargeted attacks, they can still be subverted in two ways. CRF-RNN is effectively FCN8s with an appended mean-field layer. Fig. 9b

shows, that adversarial examples generated via FGSM from FCN8s (“unary” potentials) are more effective on CRF-RNN than attacks from the output layer of CRF-RNN.

Also, targeted attacks with FGSM II and Iterative FGSM II are more effective since the label used to compute the loss for generating the adversarial example is not the network’s (highly confident) prediction but rather the least likely label. Consequently, the loss is high and there is a strong gradient signal from which to compute the adversarial example. Fig. 9c shows that CRF-RNN and FCN8s barely differ in their adversarial robustness to targeted attacks.

Finally, Fig. 10 shows that the same observations hold on the DeepLab v2 network, where the DenseCRF model is used as post-processing, and is not part of the neural network. This confirms that end-to-end training of the CRF, as done in CRF-RNN [91], does not influence adversarial robustness.

8.3 Discussion

The smoothing effect of CRFs, perhaps counter-intuitively, has no impact on the adversarial robustness of a DNN. However, mean-field inference produces confident marginals, making untargeted attacks less effective since they rely on the gradient of the final loss with respect to the prediction. Black-box attacks generated from models without a CRF transfer well to networks with a CRF, and are actually more effective. This is the case for both CRFs trained end-to-end [91] and used as post-processing [22], as shown in the supplementary. Finally, CRFs confer no robustness to untargeted attacks. Our investigation of the CRF also underlines the importance of testing thoroughly with black-box attacks and multiple attack algorithms, which is not the case for numerous proposed defenses [24], [37], [38], [72].

9 CONCLUSION

We have presented what to our knowledge is the first rigorous evaluation of the robustness of semantic segmentation models to adversarial attacks. We believe our main observations will facilitate future efforts to understand and defend against these attacks without compromising accuracy:

Networks with *residual connections* are inherently more robust than chain-like networks. This extends to the case of models with very few parameters, contrary to the prior observations of [53], [61] (we stress that these were however made in a different context, as they did not consider different network architectures, but only varied the number of

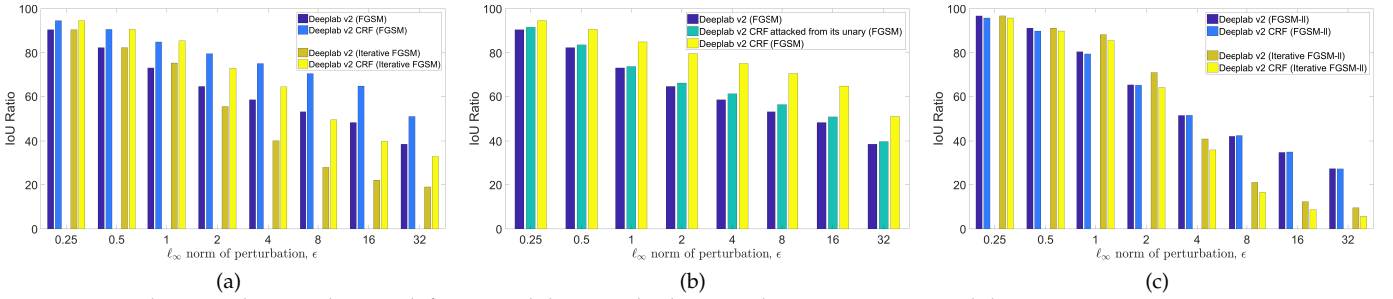


Fig. 10: Similar trends are observed for Deeplab v2, which uses the DenseCRF model as post-processing, as CRF-RNN (Fig. 9) which integrates the CRF as part of the deep network. (a) On untargeted attacks, Deeplab v2 with a CRF is noticeably more robust than just the Deeplab v2 network. (b) Attacks created 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.

filters per DNN layer). *Multiscale* processing makes CNNs more robust since adversarial inputs are not as malignant when processed at a different scale from which they were generated at, probably as CNNs are not invariant to scale. Using other *input transformations* that CNNs are not invariant make them markedly more robust to transformed adversarial examples but only when the attack generation does not take knowledge of these input transformations into account. This holds even when the input transformations are randomised. However, when this knowledge is taken into account during attack generation, only marginal improvements in robustness are observed. The fact that adversarial attacks generated to target particular input transformation do not always transfer well to other input transformations also suggests that producing physical adversarial attacks in varying environmental conditions is difficult.

Mean-field inference for Dense CRFs, which increases the confidence of predictions confers robustness to untargeted attacks, as it naturally performs “gradient masking” [70], [72]. There are no robustness benefits from the smoothness priors enforced by the DenseCRF model.

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. As the most accurate network on clean inputs is not necessarily the most robust network, we recommend evaluating robustness to a variety of adversarial attacks as done in this paper to find the best combination of accuracy and robustness before deploying models in practice. We also emphasize that it is crucial to evaluate proposed defenses judiciously, *e.g.* using the white-box attacks which exploit knowledge of the proposed defense to assess the real efficacy of such a defense.

Adversarial attacks are arguably the greatest challenge affecting DNNs. The recent interest of our field into this phenomenon is only the start of an important longer-term effort, and we should also study the influence of other factors such as training regimes and attacks tailored to evaluation metrics. In this paper, we have made numerous observations and raised questions that will aid future work in understanding adversarial examples and developing more effective defenses.

ACKNOWLEDGMENTS

This work was supported by the EPSRC, Clarendon Fund, ERC grant ERC-2012-AdG 321162-HELIOS, EPSRC grant Seebibyte EP/M013774/1 and EPSRC/MURI grant EP/N019474/1. We would also like to acknowledge the Royal Academy of Engineering and FiveAI.

REFERENCES

- [1] A. Arnab, S. Jayasumana, S. Zheng, and P. H. S. Torr. Higher order conditional random fields in deep neural networks. In *ECCV*, 2016.
- [2] A. Arnab, O. Miksik, and P. H. Torr. On the robustness of semantic segmentation models to adversarial attacks. In *CVPR*, 2018.
- [3] A. Arnab, S. Zheng, S. Jayasumana, B. Romera-Paredes, M. Larsson, A. Kirillov, B. Savchynskyy, C. Rother, F. Kahl, and P. H. S. Torr. Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction. *IEEE Signal Processing Magazine*, 35(1):37–52, Jan 2018.
- [4] A. Athalye and N. Carlini. On the robustness of the cvpr 2018 white-box adversarial example defenses. In *arXiv preprint arXiv:1804.03286*, 2018.
- [5] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.
- [6] A. Athalye and I. Sutskever. Synthesizing robust adversarial examples. In *arXiv preprint arXiv:1707.07397v1*, 2017.
- [7] V. Badrinarayanan, A. Handa, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *CoRR*, abs/1505.07293, 2015.
- [8] H. G. Barrow and J. Tenenbaum. Interpreting line drawings as three-dimensional surfaces, 1981.
- [9] Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. In *arXiv preprint arXiv:1308.3432*, 2013.
- [10] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.
- [11] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *ICML*, 2012.
- [12] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. Adversarial patch. In *arXiv preprint arXiv:1712.09665*, 2017.
- [13] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *ICLR*, 2018.
- [14] R. Bunel, I. Turkaslan, P. H. Torr, P. Kohli, and M. P. Kumar. Piecewise linear neural network verification: A comparative study. In *arXiv preprint arXiv:1711.00455*, 2017.
- [15] P. Carbonetto and N. D. Freitas. Conditional mean field. In *NIPS*, 2007.
- [16] N. Carlini and D. Wagner. Defensive distillation is not robust to adversarial examples. In *arXiv preprint arXiv:1607.04311v1*, 2016.

- [17] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *arXiv preprint arXiv:1705.07263v1*, 2017.
- [18] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017.
- [19] K. Chalupka, P. Perona, and F. Eberhardt. Visual causal feature learning. In *UAI*, 2015.
- [20] S. Chandra and I. Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *ECCV*, 2016.
- [21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *ICLR*, 2015.
- [22] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915v2*, 2016.
- [23] M. Cisse, Y. Adi, N. Neverova, and J. Keshet. Houdini: Fooling deep structured prediction models. In *NIPS*, 2017.
- [24] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval networks: Improving robustness to adversarial examples. In *ICML*, 2017.
- [25] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.
- [26] E. D. Cubuk, B. Zoph, S. S. Schoenholz, and Q. V. Le. Intriguing properties of adversarial examples. In *arXiv preprint arXiv:1711.02846*, 2017.
- [27] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.
- [28] N. Dalvi, P. Domingos, S. Sanghai, D. Verma, et al. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108. ACM, 2004.
- [29] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy. A study of the effect of jpg compression on adversarial images. In *arXiv preprint arXiv:1608.00853v1*, 2016.
- [30] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017.
- [31] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.
- [32] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song. Robust physical-world attacks on machine learning models. In *arXiv preprint arXiv:1707.08945v3*, 2017.
- [33] A. Fawzi and P. Frossard. Manitest: Are classifiers really invariant? In *BMVC*, 2015.
- [34] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner. Detecting adversarial samples from artifacts. In *arXiv preprint arXiv:1703.00410v2*, 2017.
- [35] V. Fischer, M. C. Kumar, J. H. Metzen, and T. Brox. Adversarial examples for semantic image segmentation. In *ICLR Workshop*, 2017.
- [36] D. A. Forsyth, J. Malik, M. M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler. *Finding pictures of objects in large collections of images*. Springer, 1996.
- [37] J. Gao, B. Wang, and Y. Qi. Deepmask: Masking dnn models for robustness against adversarial samples. In *ICLR Workshop*, 2017.
- [38] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.
- [39] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. On the (statistical) detection of adversarial examples. In *arXiv preprint arXiv:1702.06280v1*, 2017.
- [40] S. Gu and L. Rigazio. Towards deep neural network architectures robust to adversarial examples. In *ICLR Workshop*, 2015.
- [41] C. Guo, M. Rana, M. Cisse, and L. van der Maaten. Countering adversarial images using input transformations. In *ICLR*, 2018.
- [42] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.
- [43] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [44] W. He, J. Wei, X. Chen, N. Carlini, and D. Song. Adversarial example defenses: Ensembles of weak defenses are not strong. In *arXiv preprint arXiv:1706.04701v1*, 2017.
- [45] J. F. Henriques and A. Vedaldi. Warped convolutions: Efficient invariance to spatial transformations. In *ICML*, 2017.
- [46] J. Janai, F. Güney, A. Behl, and A. Geiger. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. In *arXiv preprint arXiv:1704.05519v1*, 2017.
- [47] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pages 97–117. Springer, 2017.
- [48] A. Kerckhoffs. La cryptographie militaire. *Journal des sciences militaires*, 9:5–83, 1883.
- [49] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *ICML*, 2017.
- [50] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, 2011.
- [51] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [52] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *ICLR Workshop*, 2017.
- [53] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In *ICLR*, 2017.
- [54] F. Liao, M. Liang, Y. Dong, T. Pang, J. Zhu, and X. Hu. Defense against adversarial attacks using high-level representation guided denoiser. In *CVPR*, 2018.
- [55] G. Lin, C. Shen, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016.
- [56] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [57] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.
- [58] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
- [59] J. Lu, H. Sibai, E. Fabry, and D. Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles. In *CVPR Workshop*, 2017.
- [60] J. Lu, H. Sibai, E. Fabry, and D. Forsyth. Standard detectors aren't (currently) fooled by physical adversarial stop signs. In *arXiv preprint arXiv:1710.03337v1*, 2017.
- [61] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *arXiv preprint arXiv:1706.06083v2*, 2017.
- [62] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. In *ICLR*, 2017.
- [63] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. In *ICCV*, 2017.
- [64] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. In *ICCV*, 2017.
- [65] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *CVPR*, 2017.
- [66] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016.
- [67] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [68] N. Narodytska and S. P. Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. In *CVPRW*, 2017.
- [69] N. Papernot, P. McDaniel, and I. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. In *arXiv preprint arXiv:1605.07277v1*, 2016.
- [70] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, 2017.
- [71] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P)*, 2016 IEEE European Symposium on, 2016.
- [72] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, 2016.
- [73] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. In *arXiv preprint arXiv:1606.02147v1*, 2016.
- [74] B. Pepik, R. Benenson, T. Ritschel, and B. Schiele. What is holding back convnets for detection? In *German Conference on Pattern Recognition*, pages 517–528. Springer, 2015.
- [75] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer. Deflecting adversarial attacks with pixel deflection. In *CVPR*, 2018.
- [76] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security*, 2016.

- [77] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [78] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *ICLR*, 2018.
- [79] J. Su, D. V. Vargas, and S. Kouichi. One pixel attack for fooling deep neural networks. In *arXiv preprint arXiv:1710.08864*, 2017.
- [80] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.
- [81] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.
- [82] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In *arXiv preprint arXiv:1705.07204v2*, 2017.
- [83] J. Uesato, B. O’Donoghue, A. v. d. Oord, and P. Kohli. Adversarial risk and the dangers of evaluating against weak attacks. In *ICML*, 2018.
- [84] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille. Mitigating adversarial effects through randomization. In *ICLR*, 2018.
- [85] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, 2017.
- [86] W. Xu, D. Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *arXiv preprint arXiv:1704.01155v1*, 2017.
- [87] X. Xu, X. Chen, C. Liu, A. Rohrbach, T. Darell, and D. Song. Can you fool ai with adversarial examples on a visual turing test? In *arXiv preprint arXiv:1709.08693v1*, 2017.
- [88] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.
- [89] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. Icnet for real-time semantic segmentation on high-resolution images. In *arXiv preprint arXiv:1704.08545v1*, 2017.
- [90] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017.
- [91] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.

Anurag Arnab is a DPhil (PhD) student at the University of Oxford, supervised by Professor Philip Torr.

Ondrej Miksik received his DPhil degree from the University of Oxford.

Philip H.S Torr received his DPhil degree from the University of Oxford and is now a professor there.