On the Robustness of Semantic Segmentation Models to Adversarial Attacks

Anurag Arnab[1] Ondrej Miksik[1,2] Philip H.S. Torr[1]
[1]University of Oxford [2]Emotech Labs
{anurag.arnab, ondrej.miksik, philip.torr}@eng.ox.ac.uk

Abstract

Deep Neural Networks (DNNs) have been demonstrated to perform exceptionally well 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 complex 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 and multiscale processing 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 which segmentation models should currently be preferred in safety-critical applications due to their inherent robustness.

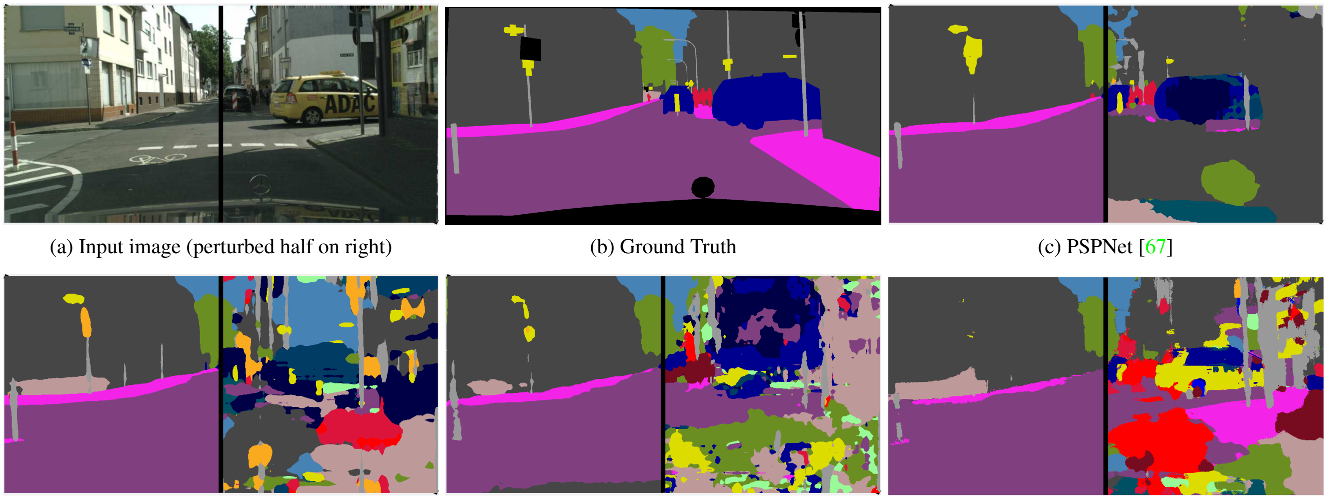
1. Introduction

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 [59]. This raises doubts about DNNs being used in safety-critical applications such as driverless vehicles [36] or medical diagnosis [21] 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 the possibility of malicious agents attacking systems that use neural networks [40, 53, 57, 23]. 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 [29, 41, 55, 48]. However, these defenses are not universal; they have frequently been found to be vulnerable to other types of attacks [11, 9, 10, 34] and/or come at the cost of performance penalties on clean inputs [12, 31, 48]. 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 CIFAR10 [48, 31, 55]. 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 [5]. This has also been witnessed by the fact that state-of-the-art semantic segmentation models are typically based on standard image classification architectures [39, 58, 33], extended by additional components such as dilated convolutions [14, 65], specialised pooling [15, 67], skip-connections [45, 7], Conditional Random Fields (CRFs) [68, 1] and/or multiscale processing [15, 13] whose impact on the robustness has never been thoroughly studied.

First, we analyse the robustness of various DNN ar-



(a) Input image (perturbed half on right)

(b) Ground Truth

(c) PSPNet [67]

(d) DilatedNet [65]

(e) ICNet [66]

(f) CRF-RNN [68]

Figure 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 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.

chitectures to adversarial examples and show that the Deeplab v2 network [15] is significantly more robust than approaches which achieve better prediction scores on public benchmarks [67]. Second, we show that adversarial examples are less effective when processed at different scales. Furthermore, multiscale networks are more robust to multiple different attacks and white-box attacks on them produce more transferable perturbations. Third, we show that structured prediction models have a similar effect as "gradient-masking" defense strategies [54, 55]. 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. Our fourth contribution shows that some widely accepted observations about robustness and model size or iterative attacks, which were made in the context of image classification [41, 48] do not transfer to semantic segmentation and different, real-world datasets. Finally, in contrast to the prior art [41, 44], our experiments are carried out on two large-scale, real-world datasets and (most of) our observations remain consistent across them. 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 neural network to change its original prediction, when added to the original input $\mathbf{x}$. For a neural network f parametrised by θ that maps $\mathbf{x} \in \mathbb{R}^m$ to y, a target class from $\mathcal{L} = \{1, 2, \ldots, L\}$, Szegedy *et al.* [59] defined an adversarial perturbation $\mathbf{r}$ as the solution to the optimisation problem

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

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, [59] added an additional term to the objective based on the loss function used to train the network

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

Here, L is the loss function between the network prediction and desired target, and c is a positive scalar. Szegedy *et al.* [59] solved this using L-BFGS, and [11] and [16] 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) [29]. 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 ll [41]. This single-step attack encourages the network to classify the adversarial example as y_t by assigning

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

We follow the convention of choosing the target class as the least likely class predicted by the network [41].

Iterative FGSM [41, 48]. 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

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

The 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 [48], with step-size α, into an ℓ_∞ ball of radius ϵ around the input $\mathbf{x}$.

Iterative FGSM ll [41]. This is a stronger version of FGSM ll. This attack sets the target to be the least likely class predicted by the network, y_{ll}, in each iteration

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

The aforementioned attacks were all proposed in the context of image classification, but they have been adapted to the problems of semantic segmentation [26, 16], object detection [62] and visual question answering [64].

3. Adversarial Defenses and Evaluations

Liu *et al*. [44] 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 [59, 54, 50, 51]. Kurakin *et al*. [41], contrastingly, studied the adversarial training defense, which generates adversarial examples online and adds them into the training set [29, 48, 60]. 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*. [48], 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*. [60] 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 [49, 30, 25, 63] have also all been subverted [10, 34, 9].

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 [41], and GoogleNet, VGG and ResNet in [44]). 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 [41, 44]. 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.

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 will publicly release our raw experimental data, evaluation code and models to aid reproducibility.

Datasets. We use the Pascal VOC [22] and Cityscapes [18] validation sets, the two most widely used semantic segmentation benchmarks. Pascal VOC consists of internet-images labelled with 21 different classes. The reduced validation [68, 45] set contains 346 images, and the training set has about 70000 images when combined with additional annotations from [32] and [43]. 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 [68], Dilated-Net [65], PSPNet [67] on Cityscapes, ICNet [67] and Seg-Net [4]. We retrained FCN [45] and E-Net [56], as well as Deeplab v2 [15] and PSPNet for VOC as the public models are trained with the validation set. Our selection of networks are based on both VGG [58] and ResNet [33] 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 [65, 15], skip-connections [45], specialised pooling [67, 15], encoder-decoder architecture [4, 56], multiscale processing [15] and CRFs [68]. 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 [68] or multiscale averaging [15] as network layers whilst training.

Adversarial attacks. We use the FGSM, FGSM ll, Iterative FGSM and Iterative FGSM ll attacks described in Sec. 2. Following [41], we set the number of iterations of iterative attacks to $\min(\epsilon + 4, \lceil 1.25\epsilon \rceil)$ and step-size $\alpha = 1$ meaning that the value of each pixel is changed by 1 every iteration. The Iterative FGSM (untargeted) and FGSM ll (targeted) attacks are only reported in the supplementary as we observed similar trends on FGSM and Iterative FGSM ll. 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 is 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 [22, 18]. However, as the accuracy of each model varies, we adapt the relative metric used by [41] 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

We evaluate the robustness of different architectures and show how our observations regarding model capacity and single-step attacks do not corroborate with some previous findings in the context of image classification [41, 48]. Additionally, our results also support why JPEG compression as a pre-processing step mitigates small perturbations [20].

5.1. The robustness of different networks

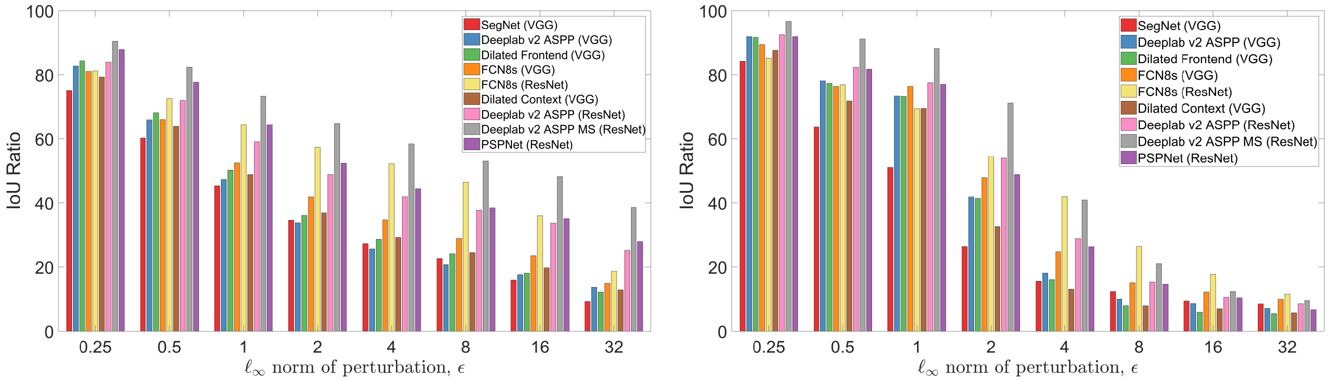
Fig. 2 shows the robustness of several state-of-the-art models on the VOC dataset. In general, ResNet-based models not only achieve higher accuracy on clean inputs but are also more robust to adversarial inputs. This is particularly the case for the single-step FGSM attack (Fig. 2a). On the more effective Iterative FGSM ll 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 [41] (Sec. 6.2). Thus, they are less useful in practical, black-box attacks.

In particular, we have evaluated the FCN8s [45] and Deeplab-v2 with ASPP [15] models based on the popular VGG-16 [58] and ResNet-101 [33] 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 [67] is actually less robust than Deeplab v2 with Multiscale ASPP [15]: 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 ll 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.* [48] and Kurakin *et al.* [41] have studied the effect of model capacity on adversarial robustness by changing the number of filters at each DNN layer, since they used the parameter count as a proxy for model capacity. Madry *et al.* [48] 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 would have serious safety implications on deployment of lightweight models, typically required by embedded platforms. Instead, we analyse different network structures and show in Fig. 3 that lightweight networks such as E-Net [56] (only 1.5 MB) and IC-Net [66] (only 30.1 MB) are affected by adversarial examples similarly as Dilated-Net [65] 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-ll 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



(a) Untargeted FGSM attack

(b) Targeted Iterative FGSM ll. attack

Figure 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 ll 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.

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, 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 ll attacks are significantly more effective on Cityscapes than on Pascal VOC. The IoU ratio for FGSM at $\epsilon = 32$ for PSPNet and Dilated Context is 2.5% and 2.8%, respectively, on Cityscapes. On Pascal VOC, it is substantially higher at 27.9% and 12.2%. Single-step methods (which only search in a 1-D subspace in the space of images) also appear to outperform iterative methods for high ϵ values on Cityscapes. In contrast, iterative attacks appear about as effective on Cityscapes as on Pascal VOC, when using the same hyperparameters as [41].

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 [27]. The effect of the training set on adversarial attacks has not been considered before, and most prior work used MNIST [59, 29, 48] or ImageNet [41, 60, 44]. However, [6] and [37], 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 ll attack, which shows the same trend as FGSM, are in the supplementary.

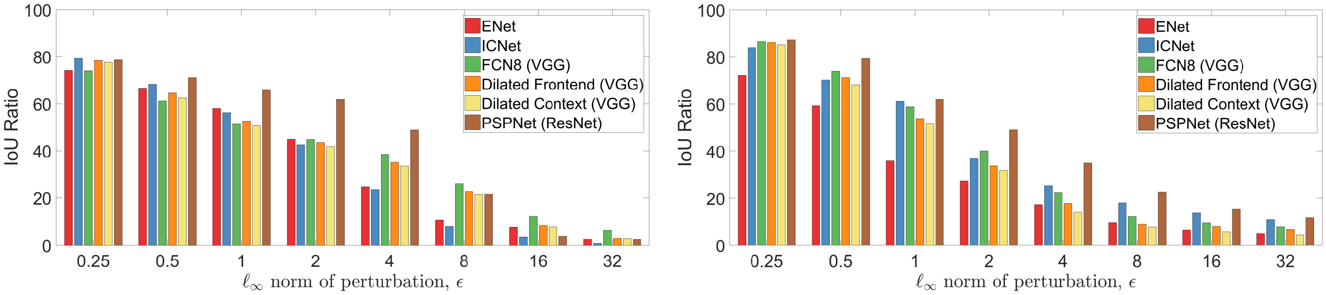
5.4. Imperceptible perturbations

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 9% relative IoU for each attack. The observation of [20], 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. Discussion

We have showed that models with residual connections (ResNet, E-Net, ICNet) are inherently more robust than 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 [48] and [41] who simply increased the number of filters at each layer. The most robust model was Deeplab v2 with Multiscale ASPP, outperforming the current state-of-the-art PSPNet [67], 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. This was unexpected, given that single-step methods only search in a one-dimensional subspace, and raises 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.



(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.* [48], we observe that lightweight networks such as E-Net [56] and ICNet [66] are often about as robust as Dilated-Net [65] (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.

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 [15]. 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 [14, 19, 42, 67].

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 [24, 35]. 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.

6.2. The transferability of adversarial examples at different scales

Table 1 shows results for the FGSM and Iterative FGSM ll 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 constrast to Iterative FGSM ll, FGSM attacks transfer well to other networks, which confirms the observations [41] 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 ll 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 in-

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.

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**

puts 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. Transformations of adversarial examples

Adversarial examples do not transfer well across different scales and transformations, as noted by Lu *et al.* [46]. The authors created adversarial traffic signs after capturing images of them from 0.5m and 1.5m away. Whilst the printed image taken from 0.5m fooled an object detector viewing the image from 0.5m, it did not when viewed from 1.5m and vice versa. This result is corroborated by Tab.1 which shows adversarial examples transfer poorly across different scales. As CNNs are not invariant to many classes of transformations (including scale) [24], adversarial examples undergoing them will not be as malicious since the activations of the network change greatly compared to the original input. Whilst we have shown that networks are more vulnerable to black-box perturbations generated from multiple scales, there may be other transformations which are even more difficult to model for the attacks. This effectively makes it more challenging to produce physical adversarial examples in the real world [47] which can be processed from a wide range of viewpoints and camera distortions.

6.5. Relation to other defenses

Our observations relate to the "random resizing" defense of [61] 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. We hypothesise that this defense could be defeated by creating adversarial attacks from multiple scales, as done in this work and concurrently in [3].

7. Effect of CRFs on Adversarial Robustness

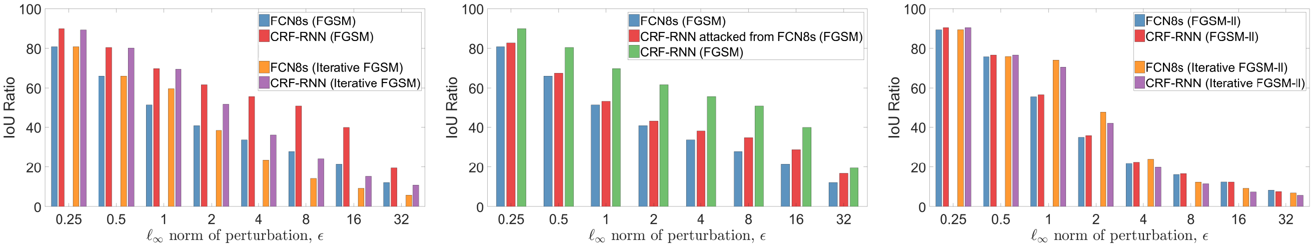
Conditional Random Fields (CRFs) are commonly used in semantic segmentation to enforce structural constraints [2]. The most common formulation is DenseCRF [38], which encourages nearby (in terms of position or appearance) pixels to take on the same label and hence prefers smooth labelling. This is done by a pairwise potential function, 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 [68]. This approach formulates mean-field inference of DenseCRF as an RNN which is appended to the FCN8s network [45], enabling end-to-end training. We also report in the supplementary material similar results for DeepLab v2, which performs mean-field inference as a post-processing step.

7.1. CRFs confer robustness to untargeted attacks

Fig. 4a 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. 4a, 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



(a) (b) (c)

Figure 4: (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.

mean-field inference tends to produce overconfident predictions has also been noted previously by [52] and [8].

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 [55] 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 [55] on accuracy is unknown, as it was only tested on the saturated MNIST and CIFAR10 datasets.

7.2. 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. 4b 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 ll and Iterative FGSM ll 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. 4c shows that CRF-RNN and FCN8s barely differ in their adversarial robustness to targeted attacks.

7.3. 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 [68] and used as post-processing [15], 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 [17, 28, 29, 55].

8. 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 [41, 48]. *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. The fact that CNNs are not invariant to many classes of transformations also makes producing physical adversarial attacks more difficult. Note, however, that multiscale perturbations also transfer better to other models. *Mean-field inference for Dense CRFs*, which increases the confidence of predictions confers robustness to untargeted attacks, as it naturally performs "gradient masking" [54, 55].

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.

Adversarial attacks are arguably the greatest challenge affecting DNNs. The recent interest 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.

We thank Qizhu Li and Bernardino Romera-Paredes for valuable input. 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.

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. 1

[2] 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, 2018. 7

[3] A. Athalye and I. Sutskever. Synthesizing robust adversarial examples. In *arXiv preprint arXiv:1707.07397v1*, 2017. 7

[4] 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. 4

[5] H. G. Barrow and J. Tenenbaum. Interpreting line drawings as three-dimensional surfaces, 1981. 1

[6] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *ICML*, 2012. 5

[7] P. Bilinski and V. Prisacariu. Dense Decoder Shortcut Connections for Single-Pass Semantic Segmentation. In *CVPR*, 2018. 1

[8] P. Carbonetto and N. D. Freitas. Conditional mean field. In *NIPS*, 2007. 8

[9] N. Carlini and D. Wagner. Defensive distillation is not robust to adversarial examples. In *arXiv preprint arXiv:1607.04311v1*, 2016. 1, 3

[10] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *arXiv preprint arXiv:1705.07263v1*, 2017. 1, 3

[11] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017. 1, 2

[12] K. Chalupka, P. Perona, and F. Eberhardt. Visual causal feature learning. In *UAI*, 2015. 1

[13] S. Chandra and I. Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *ECCV*, 2016. 1

[14] 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. 1, 6

[15] 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. 1, 2, 4, 6, 8

[16] M. Cisse, Y. Adi, N. Neverova, and J. Keshet. Houdini: Fooling deep structured prediction models. In *NIPS*, 2017. 2, 3

[17] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval networks: Improving robustness to adversarial examples. In *ICML*, 2017. 8

[18] 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. 3, 4

[19] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015. 6

[20] 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. 4, 5

[21] 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. 1

[22] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 3, 4

[23] 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. 1

[24] A. Fawzi and P. Frossard. Manitest: Are classifiers really invariant? In *BMVC*, 2015. 6, 7

[25] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner. Detecting adversarial samples from artifacts. In *arXiv preprint arXiv:1703.00410v2*, 2017. 3

[26] V. Fischer, M. C. Kumar, J. H. Metzen, and T. Brox. Adversarial examples for semantic image segmentation. In *ICLR Workshop*, 2017. 3

[27] 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. 5

[28] J. Gao, B. Wang, and Y. Qi. Deepmask: Masking dnn models for robustness against adversarial samples. In *ICLR Workshop*, 2017. 8

[29] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1, 2, 3, 5, 8

[30] 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. 3

[31] S. Gu and L. Rigazio. Towards deep neural network architectures robust to adversarial examples. In *ICLR Workshop*, 2015. 1

[32] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 3

[33] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 4

[34] 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. 1, 3

[35] J. F. Henriques and A. Vedaldi. Warped convolutions: Efficient invariance to spatial transformations. In *ICML*, 2017. 6

[36] 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. 1

[37] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *ICML*, 2017. 5

[38] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, 2011. 7

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012. 1

[40] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *ICLR Workshop*, 2017. 1

[41] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In *ICLR*, 2017. 1, 2, 3, 4, 5, 6, 8

[42] G. Lin, C. Shen, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016. 6

[43] 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. 3

[44] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017. 2, 3, 5

[45] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 3, 4, 7

[46] 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. 7

[47] 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. 7

[48] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1, 2, 3, 4, 5, 6, 8

[49] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. In *ICLR*, 2017. 3

[50] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. In *ICCV*, 2017. 3

[51] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *CVPR*, 2017. 3

[52] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 8

[53] 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. 1

[54] 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. 2, 3, 8

[55] 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. 1, 2, 8

[56] 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. 4, 6

[57] 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. 1

[58] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1, 4

[59] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 1, 2, 3, 5

[60] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In *arXiv preprint arXiv:1705.07204v2*, 2017. 3, 5

[61] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille. Mitigating adversarial effects through randomization. In *ICLR*, 2018. 7

[62] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, 2017. 3

[63] W. Xu, D. Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *arXiv preprint arXiv:1704.01155v1*, 2017. 3

[64] 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. 3

[65] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 1, 2, 4, 6

[66] 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. 2, 4, 6

[67] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 2, 4, 5, 6

[68] 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. 1, 2, 3, 4, 7, 8