Indirect Local Attacks for Context-aware Semantic Segmentation Networks

Krishna Kanth Nakka[1] and Mathieu Salzmann[1,2]

[1]CVLab, EPFL, Switzerland [2]ClearSpace, Switzerland
{krishna.nakka, mathieu.salzmann}@epfl.ch

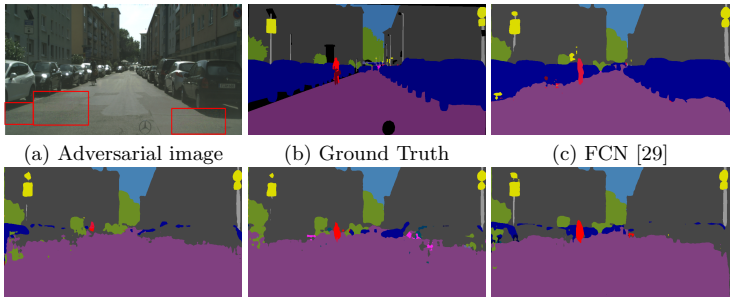
Abstract. Recently, deep networks have achieved impressive semantic segmentation performance, in particular thanks to their use of larger contextual information. In this paper, we show that the resulting networks are sensitive not only to global adversarial attacks, where perturbations affect the entire input image, but also to indirect local attacks, where the perturbations are confined to a small image region that does not overlap with the area that the attacker aims to fool. To this end, we introduce an indirect attack strategy, namely adaptive local attacks, aiming to find the best image location to perturb, while preserving the labels at this location and producing a realistic-looking segmentation map. Furthermore, we propose attack detection techniques both at the global image level and to obtain a pixel-wise localization of the fooled regions. Our results are unsettling: Because they exploit a larger context, more accurate semantic segmentation networks are more sensitive to indirect local attacks. We believe that our comprehensive analysis will motivate the community to design architectures with contextual dependencies that do not trade off robustness for accuracy.

Keywords: Adversarial Attacks, Semantic Segmentation

1 Introduction

Deep Neural Networks (DNNs) are highly expressive models and achieve state-of-the-art performance in many computer vision tasks. In particular, the powerful backbones originally developed for image recognition have now be recycled for semantic segmentation, via the development of fully convolutional networks (FCNs) [29]. The success of these initial FCNs, however, was impeded by their limited understanding of surrounding context. As such, recent techniques have focused on exploiting contextual information via dilated convolutions [50], pooling operations [26, 53], or attention mechanisms [54, 12].

Despite this success, recent studies have shown that DNNs are vulnerable to adversarial attacks. That is, small, dedicated perturbations to the input images can make a network produce virtually arbitrarily incorrect predictions. While this has been mostly studied in the context of image recognition [35, 23, 9, 34, 39], a few recent works have nonetheless discussed such adversarial attacks for semantic segmentation [49, 2, 18]. These methods, however, remain limited to



(a) Adversarial image (b) Ground Truth (c) FCN [29]

(d) PSPNet [53] (e) PSANet [54] (f) DANet [12]

Fig. 1: **Indirect Local Attacks.** An adversarial input image **(a)** is attacked with an imperceptible noise in local regions, shown as red boxes, to fool the dynamic objects. Such *indirect* local attacks barely affect an FCN [29] **(c)**. By contrast, modern networks that leverage context to achieve higher accuracy on clean (unattacked) images, such as PSPNet [53] **(d)**, PSANet [54] **(e)** and DANet [12] **(f)** are more strongly affected, even in regions far away from the perturbed area.

global perturbations to the entire image. Here, we argue that local attacks are more realistic, in that, in practice, they would allow an attacker to modify the physical environment to fool a network. This, in some sense, was the task addressed in [11], where stickers were placed on traffic poles so that an image recognition network would misclassify the corresponding traffic signs. In this scenario, however, the attack was directly performed on the targeted object.

In this paper, by contrast, we study the impact of *indirect local* attacks, where the perturbations are performed on regions outside the targeted objects. This, for instance, would allow one to place a sticker on a building to fool a self-driving system such that all nearby dynamic objects, such as cars and pedestrians, become mislabeled as the nearest background class. We choose this setting not only because it allows the attacker to perturb only a small region in the scene, but also because it will result in realistic-looking segmentation maps. By contrast, untargeted attacks would yield unnatural outputs, which can much more easily be detected by a defense mechanism. However, designing such targeted attacks that are effective is much more challenging than untargeted ones.

To achieve this, we first investigate the general idea of *indirect* attacks, where the perturbations can occur anywhere in the image except on the targeted objects. We then switch to the more realistic case of *localized* indirect attacks, and design a group sparsity-based strategy to confine the perturbed region to a small area outside of the targeted objects. For our attacks to remain realistic and imperceptible, we perform them without ground-truth information about the dynamic objects and in a norm-bounded setting. In addition to these indirect attacks, we evaluate the robustness of state-of-the-art networks to a single universal fixed-size local perturbation that can be learned from all training images to attack an unseen image in an untargeted manner.

The conclusions of our experiments are disturbing: In short, more accurate semantic segmentation networks are more sensitive to indirect local attacks. This is illustrated by Figure 1, where perturbing a few patches in static regions has much larger impact on the dynamic objects for the context-aware PSPNet [53], PSANet [54] and DANet [12] than for a simple FCN [29]. This, however, has to be expected, because the use of context, which improves segmentation accuracy, also increases the network's receptive field, thus allowing the perturbation to be propagated to more distant image regions.

Motivated by this unsettling sensitivity of segmentation networks to indirect local attacks, we then turn our focus to adversarial attack detection. In contrast to the only two existing works that have tackled attack detection for semantic segmentation [48, 25], we perform detection not only at the global image level, but locally at the pixel level. Specifically, we introduce an approach to localizing the regions whose predictions were affected by the attack, i.e., not the image regions that were perturbed. In an autonomous driving scenario, this would allow one to focus more directly on the potential dangers themselves, rather than on the image regions that caused them.

To summarize, our contributions are as follows. We introduce the idea of indirect local adversarial attacks for semantic segmentation networks, which better reflects potential real-world dangers. We design an adaptive, image-dependent local attack strategy to find the minimal location to perturb in the static image region. We show the vulnerability of modern networks to a universal, image-independent adversarial patch. We study the impact of context on a network's sensitivity to our indirect local attacks. We introduce a method to detect indirect local attacks at both image level and pixel level. Our code is available at `https://github.com/krishnakanthnakka/Indirectlocalattacks/`.

2 Related Work

Context in Semantic Segmentation Networks. While context has been shown to improve the results of traditional semantic segmentation methods [17, 21, 22, 13], the early deep fully-convolutonal semantic segmentation networks [29, 15] only gave each pixel a limited receptive field, thus encoding relatively local relationships. Since then, several solutions have been proposed to account for wider context. For example, UNet [42] uses contracting path to capture larger context followed by a expanding path to upsample the intermediate low-resolution representation back to the input resolution. ParseNet [26] relies on global pooling of the final convolutional features to aggregate context information. This idea was extended to using different pooling strides in PSPNet [53], so as to encode different levels of context. In [50], dilated convolutions were introduced to increase the size of the receptive field. PSANet [54] is designed so that each local feature vector is connected to all the other ones in the feature map, thus learning contextual information adaptively. EncNet [52] captures context via a separate network branch that predicts the presence of the object categories in the scene without localizing them. DANet [12] uses a dual attention mecha-

nism to attend to the most important spatial and channel locations in the final feature map. In particular, the DANet position attention module selectively aggregates the features at all positions using a weighted sum. In practice, all of these strategies to use larger contextual information have been shown to outperform simple FCNs on clean samples. Here, however, we show that this makes the resulting networks more vulnerable to indirect local adversarial attacks, even when the perturbed region covers less than 1% of the input image.

Adversarial Attacks on Semantic Segmentation: Adversarial attacks aim to perturb an input image with an imperceptible noise so as to make a DNN produce erroneous predictions. So far, the main focus of the adversarial attack literature has been image classification, for which diverse attack and defense strategies have been proposed [14, 6, 35, 23, 9, 34, 39]. In this context, it was shown that deep networks can be attacked even when one does not have access to the model weights [28, 37], that attacks can be transferred across different networks [45], and that universal perturbations that can be applied to any input image exist [32, 33, 40].

Motivated by the observations made in the context of image classification, adversarial attacks were extended to semantic segmentation. In [2], the effectiveness of attack strategies designed for classification was studied for different segmentation networks. In [49], a dense adversary generation attack was proposed, consisting of projecting the gradient in each iteration with minimal distortion. In [18], a universal perturbation was learnt using the whole image dataset. Furthermore, [4] demonstrated the existence of perturbations that are robust over chosen distributions of transformations.

None of these works, however, impose any constraints on the location of the attack in the input image. As such, the entire image is perturbed, which, while effective when the attacker has access to the image itself, would not allow one to physically modify the scene so as to fool, e.g., autonomous vehicles. This, in essence, was the task first addressed in [5], where a universal targeted patch was shown to fool a recognition system to a specific target class. Later, patch attacks were studied in a more realistic setting in [11], where it was shown that placing a small, well-engineered patch on a traffic sign was able to fool a classification network into making wrong decisions. While these works focused on classification, patch attacks have been extended to object detection [44, 27, 43, 30] and optical flow [41]. Our work differs fundamentally from these methods in the following ways. First, none of these approaches optimize the location of the patch perturbation. Second, [27, 5, 43] learn a separate perturbation for every target class, which, at test time, lets the attacker change the predictions to one class only. While this is suitable for recognition, it does not apply to our segmentation setup, where we seek to misclassify the dynamic objects as different background classes so as to produce a realistic segmentation map. Third, unlike [5, 41, 11], our perturbations are imperceptible. Finally, while the perturbations in [11, 44, 41] cover the regions that should be misclassified, and in [27, 43] affect the predictions in the perturbed region, we aim to design an attack that affects only targeted locations outside the perturbed region.

In other words,we study the impact of *indirect* local attacks, where the perturbation is outside the targeted area. This would allow one to modify static portions of the scene so as to, e.g., make dynamic objects disappear to fool the self-driving system. Furthermore, we differ from these other patch-based attacks in that we study local attacks for semantic segmentation to understand the impact of the contextual information exploited by different networks, and introduce detection strategies at both image- and pixel-level.Similarly to most of the existing literature [18, 25, 5, 11, 43, 44], we focus on the white-box setting for three reasons: (1) Developing effective defense mechanisms for semantic segmentation, which are currently lacking, requires assessing the sensitivity of semantic segmentation networks to the strongest attacks, i.e., white-box ones; (2) Recent model extraction methods [38, 46, 7] allow an attacker to obtain a close approximation of the deployed model. 3) While effective in classification [37], black-box attacks were observed to transfer poorly across semantic segmentation architectures [48], particularly in the targeted scenario. We nonetheless evaluate black-box attacks in the supplementary material.

When it comes to detecting attacks to semantic segmentation networks, only two techniques have been proposed [48, 25]. In [48], detection is achieved by checking the consistency of predictions obtained from overlapping image patches. In [25], the attacked label map is passed through a pix2pix generator [19] to resynthesize an image, which is then compared with the input image to detect the attack. In contrast to these works that need either multiple passes through the network or an auxiliary detector, we detect the attack by analyzing the internal subspaces of the segmentation network. To this end, inspired by the algorithm of [24] designed for image classification, we compute the Mahalanobis distance of the features to pre-trained class conditional distributions. In contrast to [48, 25], which study only global image-level detection, we show that our approach is applicable at both the image and the pixel level, yielding the first study on localizing the regions fooled by the attack.

3 Indirect Local Segmentation Attacks

Let us now introduce our diverse strategies to attack a semantic segmentation network. In semantic segmentation, given a clean image $\mathbf{X} \in \mathbb{R}^{W \times H \times C}$, where W, H and C are the width, height, and number of channels, respectively, a network is trained to minimize a loss function of the form $L(\mathbf{X}) = \sum_{j=1}^{W \times H} J(y_j^{true}, f(\mathbf{X})_j)$, where J is typically taken as the cross-entropy between the true label y_j^{true} and the predicted label $f(\mathbf{X})_j$ at spatial location j. In this context, an adversarial attack is carried out by optimizing for a perturbation that forces the network to output wrong labels for some (or all) of the pixels. Below, we denote by $\mathbf{F} \in \{0,1\}^{W \times H}$ the fooling mask such that $\mathbf{F}_j = 1$ if the j-th pixel location is targeted by the attacker to be misclassified and $\mathbf{F}_j = 0$ is the predicted label should be preserved. We first present our different local attack strategies, and finally introduce our attack detection technique.

3.1 Indirect Local Attacks

To study the sensitivity of segmentation networks, we propose to perform local perturbations, confined to predefined regions such as class-specific regions or patches, and to fool other regions than those perturbed. For example, in the context of automated driving, we may aim to perturb only the regions belonging to the road to fool the car regions in the output label map. This would allow one to modify the physical, static scene while targeting dynamic objects.

Formally, given a clean image $\mathbf{X} \in \mathbb{R}^{W \times H \times C}$, we aim to find an additive perturbation $\delta \in \mathbb{R}^{W \times H \times C}$ within a perturbation mask $\mathbf{M}$ that yields erroneous labels within the fooling mask $\mathbf{F}$. To achieve this, we define the perturbation mask $\mathbf{M} \in \{0,1\}^{W \times H}$ such that $\mathbf{M}_j = 1$ if the j-th pixel location can be perturbed and $\mathbf{M}_j = 0$ otherwise. Let $\mathbf{y}_j^{pred}$ be the label obtained from the clean image at pixel j. An untargeted attack can then be expressed as the solution to the optimization problem

$$\delta^* = \arg\min_{\delta} \sum_{j|\mathbf{F}_j=1} -J(\mathbf{y}_j^{pred}, f(\mathbf{X} + \mathbf{M} \odot \delta)_j) + \sum_{j|\mathbf{F}_j=0} J(\mathbf{y}_j^{pred}, f(\mathbf{X} + \mathbf{M} \odot \delta)_j) \quad (1)$$

which aims to minimize the probability of $\mathbf{y}_j^{pred}$ in the targeted regions while maximizing it in the rest of the image. By contrast, for a targeted attack whose goal is to misclassify any pixel j in the fooling region to a pre-defined label $\mathbf{y}_j^t$, we write the optimization problem

$$\delta^* = \arg\min_{\delta} \sum_{j|\mathbf{F}_j=1} J(\mathbf{y}_j^t, f(\mathbf{X} + \mathbf{M} \odot \delta)_j) + \sum_{i|\mathbf{F}_j=0} J(\mathbf{y}_j^{pred}, f(\mathbf{X} + \mathbf{M} \odot \delta)_j) \; . \quad (2)$$

We solve (1) and (2) via the iterative projected gradient descent algorithm [3] with an ℓ_p-norm perturbation budget $\|\mathbf{M} \odot \delta\|_p < \epsilon$, where $p \in \{2, \infty\}$.

Note that the formulations above allow one to achieve any local attack. To perform *indirect* local attacks, we simply define the masks $\mathbf{M}$ and $\mathbf{F}$ so that they do not intersect, i.e., $\mathbf{M} \odot \mathbf{F} = \mathbf{0}$, where $\odot$ is the element-wise product.

3.2 Adaptive Indirect Local Attacks

The attacks described in Section 3.1 assume the availability of a fixed, predefined perturbation mask $\mathbf{M}$. In practice, however, one might want to find the best location for an attack, as well as make the attack as local as possible. In this section, we introduce an approach to achieving this by enforcing structured sparsity on the perturbation mask.

To this end, we first re-write the previous attack scheme under an ℓ_2 budget as an optimization problem. Let $J_t(\mathbf{X}, \mathbf{M}, \mathbf{F}, \delta, f, \mathbf{y}^{pred}, \mathbf{y}^t)$ denote the objective function of either (1) or (2), where $\mathbf{y}^t$ can be ignored in the untargeted case. Following [6], we write an adversarial attack under an ℓ_2 budget as the solution to the optimization problem

$$\delta^* = \arg\min_{\delta} \; \lambda_1 \|\delta\|_2^2 + J_t(\mathbf{X}, \mathbf{M}, \mathbf{F}, \delta, f, \mathbf{y}^{pred}, \mathbf{y}^t) \; , \quad (3)$$

where λ_1 balances the influence of the term aiming to minimize the magnitude of the perturbation. While solving this problem, we further constrain the resulting adversarial image $\mathbf{X}+ \mathbf{M} \odot \delta$ to lie in the valid pixel range [0,1].

To identify the best location for an attack together with confining the perturbations to as small an area as possible, we divide the initial perturbation mask $\mathbf{M}$ into T non-overlapping patches. This can be achieved by defining T masks $\{\mathbf{M}_t \in \mathbb{R}^{W \times H}\}$ such that, for any s, t, with $s \neq t$, $\mathbf{M}_s \odot \mathbf{M}_t = \mathbf{0}$, and $\sum_{t=1}^{T} \mathbf{M}_t = \mathbf{M}$. Our goal then becomes that of finding a perturbation that is non-zero in the smallest number of such masks. This can be achieved by modifying (3) as

$$\delta^* = \arg\min_{\delta} \ \lambda_2 \sum_{t=1}^{T} \|\mathbf{M}_t \odot \delta\|_2 + \lambda_1 \|\delta\|_2^2 + J_t(\mathbf{X}, \mathbf{M}, \mathbf{F}, \delta, f, \mathbf{y}^{pred}, \mathbf{y}^t) \ , \qquad (4)$$

whose first term encodes an $\ell_{2,1}$ group sparsity regularizer encouraging complete groups to go to zero. Such a regularizer has been commonly used in the sparse coding literature [51, 36], and more recently in the context of deep networks for compression purposes [47, 1]. In our context, this regularizer encourages as many as possible of the $\{\mathbf{M}_t \odot \delta\}$ to go to zero, and thus confines the perturbation to a small number of regions that most effectively fool the targeted area $\mathbf{F}$. λ_2 balances the influence of this term with the other ones.

3.3 Universal Local Attacks

The strategies discussed in Sections 3.1 and 3.2 are image-specific. However, [5] showed the existence of a universal perturbation patch that can fool an image classification system to output any target class. In this section, instead of finding optimal locations for a specific image, we aim to learn a single fixed-size local perturbation that can fool any unseen image in an untargeted manner. This will allow us to understand the contextual dependencies of a fixed size universal patch on the output of modern networks. Unlike in the above-mentioned adaptive local attacks, but as in [5, 41, 28, 43], such a universal patch attack will require a larger perturbation norm. Note also that, because it is image-independent, the resulting attack will typically not be indirect. While [5] uses a different patch for different target classes, we aim to learn a single universal local perturbation that can fool all classes in the image. This will help to understand the propagation of the attack in modern networks using different long-range contextual connections. As will be shown in our experiments in Section 4.3, such modern networks are the most vulnerable to universal attacks, while their effect on FCNs is limited to the perturbed region. To find a universal perturbation effective across all images, we write the optimization problem

$$\delta^* = \arg\min_{\delta} \frac{1}{N} \sum_{i=1}^{N} J_u(\mathbf{X}^i, \mathbf{M}, \mathbf{F}^i, \delta, f, \mathbf{y}_i^{pred}) \ , \qquad (5)$$

where $J_u(\cdot)$ is the objective function for a single image, as in the optimization problem (1), N is the number of training images, $\mathbf{X}^i$ is the i-th image with

fooling mask $\mathbf{F}^i$, and the mask $\mathbf{M}$ is the global perturbation mask used for all images. In principle, $\mathbf{M}$ can be obtained by sampling patches over all possible image locations. However, we observed such a strategy to be unstable during learning. Hence, in our experiments, we confine ourselves to one or a few fixed patch positions. Note that, to give the attacker more flexibility, we take the universal attack defined in (5) to be an untargeted attack.

3.4 Adversarial Attack Detection

To understand the strength of the attacks discussed above, we introduce a detection method that can act either at the global image level or at the pixel level. The latter is particularly interesting in the case of indirect attacks, where the perturbation regions and the fooled regions are different. In this case, our goal is to localize the pixels that were fooled, which is more challenging than finding those that were perturbed, since their intensity values were not altered. To this end, we use a score based on the Mahalanobis distance defined on the intermediate feature representations. This is because, as discussed in [24, 31] in the context of image classification, the attacked samples can be better characterized in the representation space than in the output label space. Specifically, we use a set of training images to compute class-conditional Gaussian distributions, with class-specific means μ_c^ℓ and covariance $\mathbf{\Sigma}^\ell$ shared across all C classes, from the features extracted at every intermediate layer ℓ of the network within locations corresponding to class label c. We then define a confidence score for each spatial location j in layer ℓ as $C(\mathbf{X}_j^\ell) = \max_{c\in[1,C]} -\left(\mathbf{X}_j^\ell - \mu_c^\ell\right)^\top \mathbf{\Sigma}_\ell{}^{-1} \left(\mathbf{X}_j^\ell - \mu_c^\ell\right)$, where $\mathbf{X}_j^\ell$ denotes the feature vector at location j in layer ℓ. We handle the different spatial feature map sizes in different layers by resizing all of them to a fixed spatial resolution. We then concatenate the confidence scores in all layers at every spatial location and use the resulting L-dimensional vectors, with L being the number of layers, as input to a logistic regression classifier with weights $\{\alpha_\ell\}$. We then train this classifier to predict whether a pixel was fooled or not. At test time, we compute the prediction for an image location j as $\sum_\ell \alpha_\ell C(\mathbf{X}_j^\ell)$. To perform detection at the global image level, we sum over the confidence scores of all spatial positions. That is, for layer ℓ, we compute an image-level score as $C(\mathbf{X}^\ell) = \sum_j C(\mathbf{X}_j^\ell)$. We then train another logistic regression classifier using these global confidence scores as input.

4 Experiments

Datasets. In our experiments, we use the Cityscapes [8] and Pascal VOC [10] datasets, the two most popular semantic segmentation benchmarks. Specifically, for Cityscapes, we use the complete validation set, consisting of 500 images, for untargeted attacks, but use a subset of 150 images containing dynamic object instances of vehicle classes whose combined area covers at least 8% of the image for targeted attacks. This lets us focus on fooling sufficiently large regions, because reporting results on too small one may not be representative of the true

behavior of our algorithms. For Pascal VOC, we use 250 randomly selected images from the validation set because of the limited resources we have access to relative to the number of experiments we performed.

Models. We use publicly-available state-of-the-art models, namely FCN [29], DRNet [50], PSPNet [53], PSANet [54], DANet [12] on Cityscapes, and FCN [29] and PSANet [54] on PASCAL VOC. FCN, PSANet, PSPNet and DANet share the same ResNet [16] backbone network. We perform all experiments at the image resolution of 512×1024 for Cityscapes and 512×512 for PASCAL VOC.

Adversarial attacks. We use the iterative projected gradient descent (PGD) method with ℓ_∞ and ℓ_2 norm budgets, as described in Section 3. Following [2], we set the number of iterations for PGD to a maximum of 100, with an early termination criterion of 90% of attack success rate on the targeted objects. We evaluate ℓ_∞ attacks with a step size $\alpha \in \{1\text{e-}5, 1\text{e-}4, 1\text{e-}3, 5\text{e-}3\}$. For ℓ_2 attacks, we set $\alpha \in \{8\text{e-}3, 4\text{e-}2, 8\text{e-}2\}$. We set the maximum ℓ_p-norm of the perturbation ϵ to $100 \cdot \alpha$ for ℓ_∞ attacks, and to 100 for ℓ_2 attacks. For universal attacks, we use a higher ℓ_∞ ϵ bound of 0.3, with a step size $\alpha = 0.001$. We perform two types of attacks; targeted and untargeted. The untargeted attacks focus on fooling the network to move away from the predicted label. For the targeted attacks, we chose a safety-sensitive goal, and thus aim to fool the dynamic object regions to be misclassified as their (spatially) nearest background label. We do not use ground-truth information in any of the experiments but perform attacks based on the predicted labels only.

Evaluation metric. Following [18, 2, 49], we report the mean Intersection over Union (mIoU) and Attack Success Rate (ASR) computed over the entire dataset. The mIoU of FCN [29], DRNet [50], PSPNet [53], PSANet [54], and DANet [12] on clean samples at full resolution are 0.66, 0.64, 0.73, 0.72, and 0.67, respectively. For targeted attacks, we report the average ASR_t, computed as the percentage of pixels that were predicted as the target label. We additionally report the mIoU_u, which is computed between the adversarial and normal sample predictions. For untargeted attacks, we report the ASR_u, computed as the percentage of pixels that were assigned to a different class than their normal label prediction. Since, in most of our experiments, the fooling region is confined to local objects, we compute the metrics only within the fooling mask. We observed that the non-targeted regions retain their prediction label more than 98% of the time, and hence we report the metrics at non-targeted regions in the supplementary material. To evaluate detection, we report the Area under the Receiver Operating Characteristics (AUROC), both at image level, as in [48, 25], and at pixel level.

4.1 Indirect Local Attacks

Let us study the sensitivity of the networks to indirect local attacks. In this setting, we first perform a targeted attack, formalized in (2), to fool the dynamic object areas by allowing the attacker to perturb any region belonging to the static object classes. This is achieved by setting the perturbation mask $\mathbf{M}$ to 1 at all the static class pixels and the fooling mask $\mathbf{F}$ to 1 at all the dynamic class pixels. We report the mIoU_u and ASR_t metrics in Tables 1a and 1b on Cityscapes for ℓ_∞

Network	$\alpha = 0.00001$	$\alpha = 0.0001$	$\alpha = 0.001$	$\alpha = 0.005$
FCN [29]	0.64 / 5.0%	0.28 / 29%	0.13 / 55%	0.11 / 61%
PSPNet [53]	0.70 / 12%	0.05 / **85%**	0.00 / 89%	0.00 / **90%**
PSANet [54]	0.59 / **14%**	0.03 / **85%**	0.01 / **90%**	0.00 / **90%**
DANet [12]	0.80 / 5.0%	0.11 / 79%	0.01 / **90%**	0.00 / **90%**
DRN [50]	0.64 / 6.0%	0.15 / 56%	0.03 / 84%	0.02 / 86%

(a) ℓ_∞ attack

Network	$\alpha = 0.008$	$\alpha = 0.04$	$\alpha = 0.08$
FCN [29]	0.60 / 10%	0.56 / 26%	0.27 / 36%
PSPNet [53]	0.67 / **19%**	0.23 / **67%**	**0.06 / 84%**
PSANet [54]	0.59 / 14%	0.21 / 63%	0.06 / 82%
DANet [12]	0.79 / 11%	0.43 / 49%	0.13 / 79%
DRN [50]	0.63 / 10%	0.24 / 47%	0.13 / 64%

(b) ℓ_2 attack

Table 1: **Indirect attacks** on Cityscapes to fool dynamic classes while perturbing static ones. The numbers indicate $\text{mIoU}_u/\text{ASR}_t$, obtained using different step sizes α for ℓ_∞ and ℓ_2 attacks. The most robust network in each case is underlined and the most vulnerable models are highlighted in **bold**.

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

(a) ℓ_∞ attack

Network	$d = 0$	$d = 50$	$d = 100$	d = 150
FCN [29]	0.27 / 36%	0.79 / 2.0%	0.98 / 2.0%	0.99 / 1.0%
PSPNet [53]	0.06 / **84%**	0.18 / 73%	0.55 / 23%	0.99 / 0.0%
PSANet [54]	0.06 / 82%	0.10 / **75%**	0.14 / **66 %**	0.31 / **44%**
DANet [12]	0.13 / 79%	0.27 / 71%	0.67 / 26%	0.85 / 7.0%
DRN [50]	0.13 / 64%	0.44 / 17%	0.76 / 3.0%	0.95 / 0.0%

(b) ℓ_2 attack

Table 2: **Impact of local attacks** by perturbing pixels that are at least d pixels away from any dynamic class. We report $\text{mIoU}_u/\text{ASR}_t$ for different values of d.

and ℓ_2 attacks, respectively. As evidenced by the tables, FCN is more robust to such indirect attacks than the networks that leverage contextual information. In particular, PSANet, which uses long range contextual dependencies, and PSPNet are highly sensitive to these attacks.

To further understand the impact of indirect *local* attacks, we constrain the perturbation region to a subset of the static class regions. To do this in a systematic manner, we perturb the static class regions that are at least d pixels away from any dynamic object, and vary the value d. The results of this experiment using ℓ_2 and ℓ_∞ attacks are provided in Table 2. Here, we chose a step size $\alpha = 0.005$ for ℓ_∞ and $\alpha = 0.08$ for ℓ_2. Similar conclusions to those in the previous non-local scenario can be drawn: Modern networks that use larger receptive fields are extremely vulnerable to such perturbations, even when they are far away from the targeted regions. By contrast, FCN is again more robust. For example, as shown in Figure 2, while an adversarial attack occurring 100 pixels away from the nearest dynamic object has a high success rate on the context-aware networks, the FCN predictions remain accurate.

4.2 Adaptive Indirect Local Attacks

We now study the impact of our approach to adaptively finding the most sensitive context region to fool the dynamic objects. To this end, we use the group sparsity

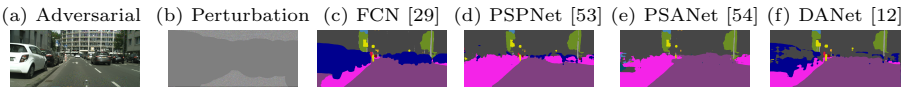


Fig. 2: **Indirect Local attack** on different networks with perturbations at least $d = 100$ pixels away from any dynamic class.

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

(a) Cityscapes

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

(b) PASCAL VOC

Table 3: **Adaptive indirect local attacks** on Cityscapes and PASCAL VOC. We report $\text{mIoU}_u/\text{ASR}_t$ for different sparsity levels S.

based optimization given in (4) and find the minimal perturbation region to fool all dynamic objects to their nearest static label. Specifically, we achieve this in two steps. First, we divide the perturbation mask $\mathbf{M}$ corresponding to all static class pixels into uniform patches of size $h \times w$, and find the most sensitive ones by solving (4) with a relatively large group sparsity weight $\lambda_2 = 100$ for Cityscapes and $\lambda_2 = 10$ for PASCAL VOC. Second, we limit the perturbation region by selecting the n patches that have the largest values $\|\mathbf{M}_t \odot \delta\|_2$, choosing n so as to achieve a given sparsity level $S \in \{75\%, 85\%, 90\%, 95\%\}$. Specifically, S is computed as the percentage of pixels that are not perturbed relative to the initial perturbation mask. We then re-optimize (4) with $\lambda_2 = 0$. In both steps, we set $\lambda_1 = 0.01$ and use the Adam optimizer [20] with a learning rate of 0.01. For Cityscapes, we use patch dimensions $h = 60$, $w = 120$, and, for PASCAL VOC, $h = 60$, $w = 60$. We clip the perturbation values below 0.005 to 0 at each iteration. This results in very local perturbation regions, active only in the most sensitive areas, as shown for PSANet in Figure 3 on Cityscapes and in Figure 4 for PASCAL VOC. As shown in Table 3, all context-aware networks are significantly affected by such perturbations, even when they are confined to small background regions. For instance, on Cityscapes, at a high sparsity level of 95%, PSANet yields an ASR_t of 44% compared to 1% for FCN. This means that, in the physical world, an attacker could add a small sticker at a static position to essentially make dynamic objects disappear from the network's view.

4.3 Universal Local Attacks

In this section, instead of considering image-dependent perturbations, we study the existence of universal local perturbations and their impact on semantic segmentation networks. In this setting, we perform untargeted *local* attacks by placing a fixed-size patch at a predetermined position. While the patch location

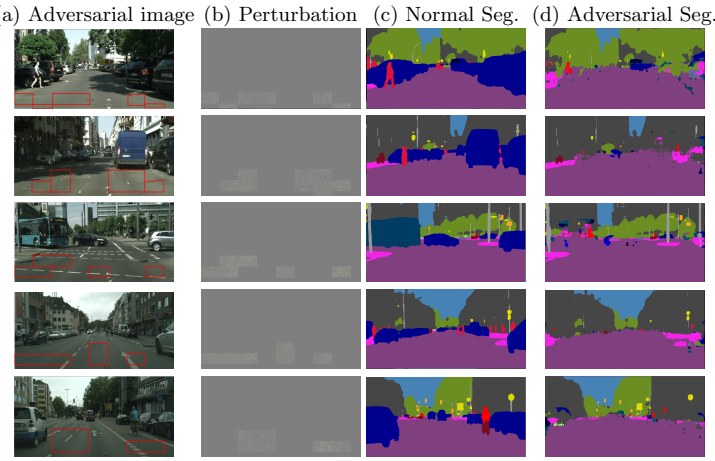


Fig. 3: **Adaptive indirect local attacks on Cityscapes with PSANet [54]**. An adversarial input image **(a)** when attacked at positions shown as red boxes with a perturbation **(b)** is misclassified within the dynamic object areas of the normal segmentation map **(c)** to result in **(d)**.

Fig. 4: **Adaptive indirect local attacks on PASCAL VOC with PSANet [54]**. An adversarial input image **(a),(e)** when attacked at positions shown as red boxes with a perturbation **(b),(f)** is misclassified within the foreground object areas of the normal segmentation map **(c), (g)** to result in **(d), (h)**, respectively.

can in principle be sampled at any location, we found learning its position to be unstable to due to the large number of possible patch locations in the entire dataset. Hence, here, we consider the scenario where the patch is located at the center of the image. We then learn a local perturbation that can fool the entire dataset of images for a given network by optimizing the objective given in (5). Specifically, the perturbation mask $\mathbf{M}$ is active only at the patch location and the fooling mask $\mathbf{F}$ at all image positions, i.e., at both static and dynamic classes. For Cityscapes, we learn the universal local perturbation using 100 images and

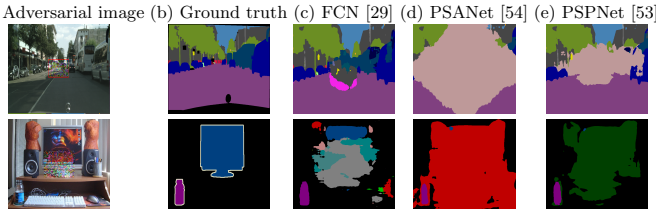


Fig. 5: **Universal local attacks** on Cityscapes and PASCAL VOC. In both datasets, the degradation in FCN [29] is limited to the attacked area, whereas for context-aware networks, such as PSPNet [53], PSANet [54], DANet [12], it extends to far away regions.

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

(a) Cityscapes

Network	51 × 51(**1.0**%)	76 × 76(**2.3**%)	102 × 102(**4.0**%)	153 × 153(**9.0**%)
FCN [29]	**0.70** / **6%**	0.70 / 7%	0.63 / 10%	0.52 / 20%
PSANet [54]	0.83 / 4%	0.76 / **8 %**	0.56 / **28%**	0.35/ **56%**

(b) PASCAL VOC

Table 4: **Universal local attacks**. We show the impact of the patch size $h \times w$ (area%) on different networks and report $\text{mIoU}_u/\text{ASR}_u$.

use the remaining 400 images for evaluation purposes. For PASCAL VOC, we perform training on 100 images and evaluate on the remaining 150 images. We use ℓ_∞ optimization with $\alpha = 0.001$ for 200 epochs on the training set. We report the results of such universal patch attacks in Tables 4a and 4b on Cityscapes and PASCAL VOC for different patch sizes. As shown in the table, PSANet and PSPNet are vulnerable to such universal attacks, even when only 2.3% of the image area is perturbed. From Figure 5, we can see that the fooling region propagates to a large area far away from the perturbed one. While these experiments study untargeted universal local attacks, we report additional results on single-class targeted universal local attacks in the supplementary material.

4.4 Attack Detection

We now turn to studying the effectiveness of the attack detection strategies described in Section 3.4. We also compare our approach to the only two detection techniques that have been proposed for semantic segmentation [48, 25]. The method in [48] uses the spatial consistency of the predictions obtained from $K = 50$ random overlapping patches of size 256×256. The one in [25] compares an image re-synthesized from the predicted labels with the input image. Both methods were designed to handle attacks that fool the entire label map, unlike our work where we aim to fool local regions. Furthermore, both methods perform detection at the image level, and thus, in contrast to ours, do not localize the fooled regions at the pixel level.

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

Table 5: Attack detection on Cityscapes with different perturbation settings.

We study detection in four perturbation settings: Global image perturbations (Global) to fool the entire image; Universal patch perturbations (UP) at a fixed location to fool the entire image; Full static (FS) class perturbations to fool the dynamic classes; Adaptive patch (AP) perturbations in the static class regions to fool the dynamic objects. As shown in Table 5, while the state-of-the-art methods [48, 25] have high Global AUROC in the first setting where the entire image is targeted, our detection strategy outperforms them by a large margin in the other scenarios. We believe this to be due to the fact that, with local attacks, the statistics obtained by studying the consistency across local patches, as in [48], are much closer to the clean image statistics. Similarly, the image re-synthesized by a pix2pix generator, as used in [25], will look much more similar to the input one in the presence of local attacks instead of global ones. For all the perturbation settings, we also report the mean percentage of pixels misclassified relative to the number of pixels in the image.

5 Conclusion

In this paper, we have studied the impact of indirect local image perturbations on the performance of modern semantic segmentation networks. We have observed that the state-of-the-art segmentation networks, such as PSANet and PSPNet, are more vulnerable to local perturbations because their use of context, which improves their accuracy on clean images, enables the perturbations to be propagated to distant image regions. As such, they can be attacked by perturbations that cover as little as 2.3% of the image area. We have then proposed a Mahalanobis distance-based detection strategy, which has proven effective for both image-level and pixel-level attack detection. Nevertheless, the performance at localizing the fooled regions in a pixel-wise manner can still be improved, which will be our goal in the future.

Acknowledgments. This work was funded in part by the Swiss National Science Foundation.

References

1. Alvarez, J.M., Salzmann, M.: Learning the number of neurons in deep networks. In: Advances in Neural Information Processing Systems. pp. 2270–2278 (2016)
2. Arnab, A., Miksik, O., Torr, P.H.: On the robustness of semantic segmentation models to adversarial attacks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 888–897 (2018)
3. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. arXiv preprint arXiv:1802.00420 (2018)
4. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: International conference on machine learning. pp. 284–293 (2018)
5. Brown, T.B., Mané, D.: Aurko roy, martín abadi, and justin gilmer. Adversarial patch. CoRR, abs/1712.09665 (2017)
6. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 39–57. IEEE (2017)
7. Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. pp. 15–26 (2017)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
9. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)
10. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge 2007 (voc2007) results (2007)
11. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1625–1634 (2018)
12. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3146–3154 (2019)
13. Gonfaus, J.M., Boix, X., Van de Weijer, J., Bagdanov, A.D., Serrat, J., Gonzalez, J.: Harmony potentials for joint classification and segmentation. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 3280–3287. IEEE (2010)
14. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
15. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 447–456 (2015)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. He, X., Zemel, R.S., Carreira-Perpiñán, M.Á.: Multiscale conditional random fields for image labeling. In: Proceedings of the 2004 IEEE Computer Society Conference

on Computer Vision and Pattern Recognition, 2004. CVPR 2004. vol. 2, pp. II–II. IEEE (2004)
18. Hendrik Metzen, J., Chaithanya Kumar, M., Brox, T., Fischer, V.: Universal adversarial perturbations against semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2755–2764 (2017)
19. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Kohli, P., Torr, P.H., et al.: Robust higher order potentials for enforcing label consistency. International Journal of Computer Vision **82**(3), 302–324 (2009)
22. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in neural information processing systems. pp. 109–117 (2011)
23. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016)
24. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Advances in Neural Information Processing Systems. pp. 7167–7177 (2018)
25. Lis, K., Nakka, K., Salzmann, M., Fua, P.: Detecting the unexpected via image resynthesis. arXiv preprint arXiv:1904.07595 (2019)
26. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579 (2015)
27. Liu, X., Yang, H., Liu, Z., Song, L., Li, H., Chen, Y.: Dpatch: An adversarial patch attack on object detectors. arXiv preprint arXiv:1806.02299 (2018)
28. Liu, Y., Chen, X., Liu, C., Song, D.: Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770 (2016)
29. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
30. Lu, J., Sibai, H., Fabry, E., Forsyth, D.: No need to worry about adversarial examples in object detection in autonomous vehicles. arXiv preprint arXiv:1707.03501 (2017)
31. Ma, X., Li, B., Wang, Y., Erfani, S.M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M.E., Bailey, J.: Characterizing adversarial subspaces using local intrinsic dimensionality. arXiv preprint arXiv:1801.02613 (2018)
32. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1765–1773 (2017)
33. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P., Soatto, S.: Analysis of universal adversarial perturbations. arXiv preprint arXiv:1705.09554 (2017)
34. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016)
35. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 427–436 (2015)
36. Nie, F., Huang, H., Cai, X., Ding, C.H.: Efficient and robust feature selection via joint l2, 1-norms minimization. In: Advances in neural information processing systems. pp. 1813–1821 (2010)

37. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277 (2016)
38. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security. pp. 506–519 (2017)
39. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P). pp. 372–387. IEEE (2016)
40. Poursaeed, O., Katsman, I., Gao, B., Belongie, S.: Generative adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4422–4431 (2018)
41. Ranjan, A., Janai, J., Geiger, A., Black, M.J.: Attacking optical flow. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2404–2413 (2019)
42. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
43. Saha, A., Subramanya, A., Patil, K., Pirsiavash, H.: Adversarial patches exploiting contextual reasoning in object detection. arXiv preprint arXiv:1910.00068 (2019)
44. Thys, S., Van Ranst, W., Goedemé, T.: Fooling automated surveillance cameras: adversarial patches to attack person detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
45. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204 (2017)
46. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction apis. In: 25th {USENIX} Security Symposium ({USENIX} Security 16). pp. 601–618 (2016)
47. Wen, W., Wu, C., Wang, Y., Chen, Y., Li, H.: Learning structured sparsity in deep neural networks. In: Advances in neural information processing systems. pp. 2074–2082 (2016)
48. Xiao, C., Deng, R., Li, B., Yu, F., Liu, M., Song, D.: Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 217–234 (2018)
49. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1369–1378 (2017)
50. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 472–480 (2017)
51. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**(1), 49–67 (2006)
52. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7151–7160 (2018)
53. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)

54. Zhao, H., Zhang, Y., Liu, S., Shi, J., Change Loy, C., Lin, D., Jia, J.: Psanet: Point-wise spatial attention network for scene parsing. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 267–283 (2018)