Adversarial Examples for Semantic Segmentation

Yuhan Xiang , Xingbin Liu , Suhang Ye , Zhiyi Chen , Wei Chen
Xiamen University,China

Abstract

It has been well demonstrated that visually imperceptible perturbations added to the natural image can successfully fool neural networks. While most of recent work has focused on image classification, we want to explore adversarial examples for semantic segmentation which is more difficult. In this paper, three white-box and one black-box attack approaches are utilized to attack segmentation neural network. We show empirically that there exist barely the perceptible perturbations which results in catastrophic predictions given by several semantic segmentation neural networks. Furthermore, we also show that adversarial examples generated from one model can be transferred to attack another one.

Introduction

Convolutional neural networks (CNNs) have become the state-of-the-art solution for a wide range of vision problems (He et al. 2015; Long, Shelhamer, and Darrell 2015; Ren et al. 2015; Szegedy et al. 2014), including object detection, visual concept discovery, semantic segmentation, boundary detection, etc. Based on powerful computational resources like modern GPUs and TPUs, 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 complex systems.

However, it is also notorious for its vulnerability which can be fooled by a human imperceptible perturbation (Carlini and Wagner 2017; Goodfellow, Shlens, and Szegedy 2015; Moosavi-Dezfooli, Fawzi, and Frossard 2016). In (Szegedy et al. 2014). In (Szegedy et al. 2014), it was shown that adding imperceptible perturbations can result in failures for image classification task. These perturbed images, called adversarial examples, are considered to fall on some areas in the large, high-dimensional feature space which are not explored in the training process.

Given that prior work on adversarial examples mostly focus on image classification task, in this paper we will explore the effect of adversarial attacks on tasks embedded a localization component, more specifically: semantic segmentation. Semantic segmentation is a widely used and important method for scene understanding which can be used for examples for automated driving, video surveillance, or robotics. With the wide-spread applicability in those domains comes the risk of being faced with an adversary trying to fool the system. Therefore, studying the adversarial attack against semantic segmentation systems becomes an important topic.

In this paper, we go one step further by generating adversarial examples for semantic segmentation and show transferability of them. Note that semantic segmentation is more difficult than classification problems, as we need to consider orders of magnitude more targeted (e.g., pixels for segmentation task). Three white-box attack methods, i.e., FGSM (Goodfellow, Shlens, and Szegedy 2015), MI-FGSM (Dong et al. 2018), PGD (Madry et al. 2019), one black-box attack,

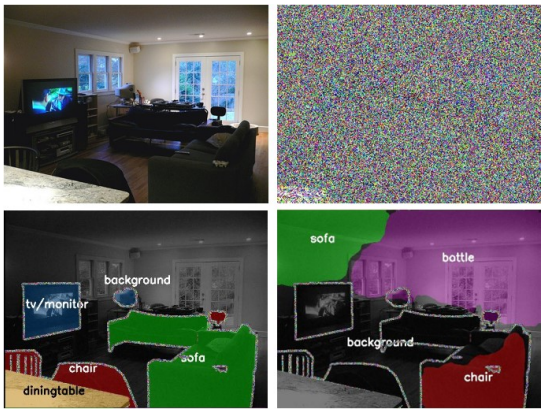

Figure 1: An adversarial example for semantic segmentation. FCN-ResNet50 is used for segmentation. Left column: the original image (top row) with the normal segmentation (bottom row). Right column: after the adversarial perturbation (top row, generated by PGD with $num_{iter} = 100$) is added to the original image, the segmentation results are completely wrong. Note that though the added perturbation can confuse semantic segmentation model, it is visually imperceptible (the maximal absolute intensity in each channel is less than 8)

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Square Attack(Andriushchenko et al. 2020), and four different segmentation networks i.e., FCN8-VGG, FCN16-VGG, FCN-Res50, FCN-Res101 are used in our work, the goal is to show how fragile current semantic segmentation models are when confronted with an adversary, and we also try to transfer perturbations generated from one model to another one to show the transferability of adversarial perturbations. Figure 1 shows an adversarial example which can confuse semantic segmentation neural network.

Related Work

Semantic Segmentation

The semantic segmentation is an essential issue in the computer vision field (Ma et al. 2017), which involves assigning a semantic category to each pixel. Jonathan Long et al. (Long, Shelhamer, and Darrell 2015) proposed Full Convolutional Networks (FCNs) for semantic segmentation in 2015, which remove the fully connected layers in classification CNN networks, becoming a pioneer of a fully convolutional architecture for dense semantic segmentation.

Based on the FCN (Long, Shelhamer, and Darrell 2015), recently many semantic segmentation approaches (Long, Shelhamer, and Darrell 2015; Paszke et al. 2016; Chen et al. 2017; Zheng et al. 2015; Zhao et al. 2017)have been proposed. Badrinarayanan V et al. (Badrinarayanan, Kendall, and Cipolla 2017) proposed the following work SegNet to introduce an encoder and decoder network, where the decoder utilizes pooling indices in the encoding layers to upsample the feature map. Ronneberger O et al. proposed UNet (Ronneberger, Fischer, and Brox 2015) adopting skip connections to combine shallow representations from the encoder and deep features from the encoder, which exploit low level feature for accurate semantic segmentation. Paszke et al. (Paszke et al. 2016) proposed ENet, real-time semantic segmentation network by exploiting separable convolution and less channels. CRF was applied as a post-processing procedure (Chen et al. 2017) or end-to-end integrated (Zheng et al. 2015) into the network to refine a segment contour. PSPNet (Zhao et al. 2017) applies pyramid pooling module to aggregate information from different scales of feature maps. Hierarchical (Tao, Sapra, and Catanzaro 2020) and ResNet (Zhang et al. 2020) have been the state-of-the-art approaches of the semantic segmentation field.

Adversarial Attack

Szegedy et al. (Szegedy et al. 2014) first showed that neural networks are vulnerable to adversarial examples, which are clean images being intentionally perturbed, e.g., by adding carefully crafted perturbation that is imperceptible to human. Various methods have been proposed to generate adversarial examples, based on the gradient (Carlini and Wagner 2017; Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2019; Dong et al. 2018), the query score (Andriushchenko et al. 2020; Huang and Zhang 2020) or the decision (Papernot et al. 2017; Cheng et al. 2018; Brendel, Rauber, and Bethge 2018; Ilyas et al. 2018), most of which mainly focus on image classification, including both white-box attack and black-box attack. In the white-box setting, both the network architecture and parameters are available to the attacker. The gradient-based methods are always useful for white-box attack and the strongest methods are almost all based on project gradient descent (PGD) (Madry et al. 2019). As for the black-box attack, the attacker only has access to the model's input and the predicted output, which is more challenging because the modification of the input must be computed without access to the loss gradient of the model. The score-based and decision-based methods work for that requiring repeated queries to the model. The state-of-the-art black-box attack is Square Attack (Andriushchenko et al. 2020), which is based on random search and square-shaped random sampling.

Adversarial Attack on Semantic Segmentation

A few studies have been conducted on the adversarial attack for semantic segmentation networks, which is also an important computer vision task and relatively more difficult. Anurag Arnab et al. (Arnab, Miksik, and Torr 2018) conducted the first systematic analysis about the effect of multiple adversarial attack methods on different semantic segmentation networks. Fisher et al. (Fischer et al. 2017) found the existence of adversarial examples in semantic segmentation and Metzen et al. (Hendrik Metzen et al. 2017) showed that universal perturbations can be made to fool semantic segmentation. Xie C et al. (Xie et al. 2017) propose an attack method named Dense Adversary Generation (DAG) to generate a group of adversarial examples for state-of-the-art segmentation and detection neural networks. Shen G et al. (Shen et al. 2019) leveraged SPADE (Spatially-adaptive denormalization) to generate effective adversarial attack in a single step, improving the attack success rate surpassing the state-of-the-art adversarial attack methods including PGD. However, due to the high cost of computation, traditional gradient-based methods such as FGSM, PGD are more efficient and thus widely used in practice.

Proposed Solution

Adversarial Attack Methods

The gradients of the loss function with respect to the input data are very common information used by adversarial attack algorithms. Generating adversarial images can be formalized as an optimization problem with constraints. Let x represent the input data and y is the corresponded label. f_θ is a DNN parametrized with θ (i.e., where the network is encoded as $f_\theta(x)$), L is the loss function which should be minimized in the standard training procedure. Under white-box setting, the gradient of the loss function with respect to input x can be easily derived:

$$\nabla_x L(f_\theta(x), y) \tag{1}$$

FGSM. Fast Gradient Sign Method (FGSM) is the simplest yet a very efficient white-box attack method. By maximizing the loss function L, adversarial examples are generated with one-step update as

$$x' = x + \epsilon \cdot sign(\nabla_x L(f_\theta(x), y)) \tag{2}$$

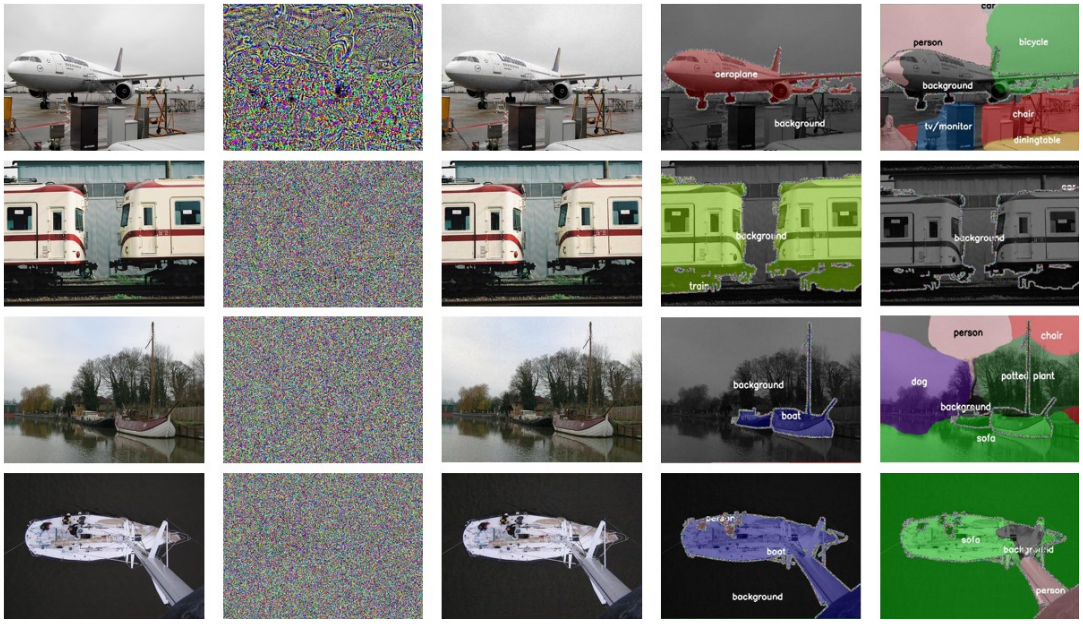

Figure 2: Four examples for semantic segmentation. These five columns, from left to right, shows the clean images, the perturbations, the adversarial images, the original segmentation result and the adversarial segmentation result, respectively. Note, the the adversarial perturbation is getting from PGD-20 attack based on FCN-ResNet50

where ϵ is the magnitude of adversarial perturbations under $l_\infty norm$ constraint.

PGD. PGD is an extended version of FGSM, by applying iterative FGSM with a small step size α and a random start point:

$$x'_0 = x, \quad x'_{t+1} = \Pi_{x+\epsilon}\{x'_t + \alpha \cdot sign(\nabla_x L\left(f_\theta\left(x'_t\right), y\right))\} \tag{3}$$

Successive variants of PGD prove that PGD is strong in white-box setting and it is usually used in adversarial defense to perform adversarial training or evaluate the robustness.

MI-FGSM. FGSM is one-step attack and get relatively lower attack success rate, while generated adversarial examples are more transferable. In contrast, the iterative method is more likely to overfit on the threat model, leading to low transferability. MI-FGSM [16] integrate momentum into the iterative FGSM to improve the transferability:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x L\left(x'_t, y\right)}{||\nabla_x L(x'_t, y)||_1} \tag{4}$$

$$x\prime_{t+1} = \Pi_{x+\epsilon} x'_t + \alpha \cdot sign\left(g_{t+1}\right) \tag{5}$$

where g_t is the accumulated gradient at iteration t, and μ is the decay factor of g_t. Like the momentum optimization formulation, the momentum help stabilize update directions and escaping from poor local maxima. MI-FGSM reaches the trade-off between the attack ability and the transferability, useful for black-box attack.

Square Attack. Square Attack is a score-based black-box attack that does not rely on the gradient information, which is based on random search scheme to sample square-shaped updates at random positions. A good initialization is important in the black-box setting, for some examples are easily modified to be adversarial. Square Attack initializes perturbations with vertical stripes of width one where the color of each stripe is sampled uniformly at random from $\{-\epsilon, \epsilon\}^c$ (c is the number of channels). Based on the observation that values in l_∞-perturbations are usually $\pm\epsilon$, the updated squares are also selected in $\{-\epsilon, \epsilon\}^c$ and the percentage of pixels p deciding the side length $\sqrt{p \cdot w^2}$ is changed as iterations grow. Finally, all elements are projected onto the l_∞-ball of radius ϵ.

Attack Semantic Segmentation

The task can be divided into two parts: build semantic segmentation models with good performance and generate adversarial examples to perform white- and black-box attack. In this paper, we perform untargeted l_∞ attack on segmentation models, fooling models segment objects to any wrong classes. For semantic segmentation, threat models are FCN8-VGG, FCN16-VGG, FCN-ResNet50 and FCN-ResNet101, where VGG is the typical VGG16 and the backbone is VGG or ResNet to extract features. The difference between FCN8, FCN16 and FCN is the upsampling operation. Segmentation models intend to achieve pixel-to-pixel classification to the input image, then output an image which is divided into several parts and corresponding label.

For adversarial attack on semantic segmentation, previous work proposed various methods and get high attack success rate. The state-of-the-art attack, AdvSPADE, leverages conditional GAN to generate adversarial examples, claiming that it improves the attack success rate significantly

Table 1: the accuracy of non-targeted adversarial attacks against four different semantic segmentation neural networks, the adversarial examples are crafted for FCN8-VGG, FCN16-VGG, FCN-Resnet50, FCN-ResNet101 respectively using FGSM, MI-FGSM, PGD attack method. Clean represent the original image of VOC dataset. Note all the attack method are white-box attacks.

	Clean	FGSM	MI-FGSM	PGD		
				20	100	200
FCN8-VGG	91.22	68.98	23.57	17.20	14.49	**14.06**
FCN16-VGG	90.99	68.22	22.13	16.11	13.45	**13.08**
FCN-Res50	93.25	71.98	16.24	16.46	**13.01**	—
FCN-Res101	94.25	74.09	22.18	21.96	**18.95**	—

and surpasses gradient-based methods. However, training GAN requires too much time, so it is not an efficient way to perform attack. Previous attack for semantic segmentation, which used gradient-based attack, only used the classic white-box attack methods, e.g., FGSM and its iterative version. Therefore, considering the efficiency and black-box attack performance, MI-FGSM and black-box Square Attack are adopted.

For white-box attack, the core of is to generate perturbations based on the gradient of loss. Semantic segmentation makes a prediction at each pixel, and the corresponding pixel-wise cross entropy loss is:

$$L_{PCE}(f(x), y) = -\frac{1}{N} \sum_{i \in N} \sum_{classes} y_i \text{LogSoftmax}(f_i(x)) \tag{6}$$

By calculating $\nabla_{x'} L_{PCE}(f(x'), y)$, we update the adversarial examples based on Eq. (2-5), maximizing L_{PCE} along the gradient direction.

For black-box attack, only logits are accessed to. The original Square Attack updates perturbations using margin-based loss $L(f(x), y) = f_y(x) - \max_{k \neq y} f_k(x)$ for untargeted attack. In image classification tasks, each image only has one label, while in segmentation, each pixel has its own label making calculation computationally intensive. To alleviate that, we adopt loss oracle to accept the update when the resulting loss is larger than the best loss so far. Pixel-wise cross entropy loss is commonly used in semantic segmentation, so it is natural to define such loss as oracle to guide the update.

Experiment

Dataset

In this project, we use PASCAL VOC2012 (Everingham et al. 2010), a benchmark in visual object recognition and detection, for training and testing. The train/val dataset has 11,530 images, which contains 27,450 ROI annotated objects and 6,929 segmentations. We should note that most Flickr images can be characterized as "snapshots", e.g., family holidays, birthdays, parties, etc. and so many objects appear only "incidentally" in images where people are the subject of the photograph. The dataset serves for object detection and semantic segmentation for it providing a large number of segmentation information under nature scene and could be an appropriate data source in our project.

Network

CNNs can be trained in end-to-end manner to accomplish semantic segmentation. We adapt ResNet and VGG, which are contemporary classification networks, into fully convolutional networks and transfer their learned representations by fine-tuning to the segmentation task, forming FCN-ResNet and FCN-VGG. Specifically, they are FCN8-VGG, FCN16-VGG, FCN-Resnet50 and FCN-ResNet101. They mainly differ in the backbone used to extract features and the architecture of the upsampling.

Different Attack Method

We report in Table 1 the accuracy of attacks against the models we consider. The adversarial examples are generated for FCN8-VGG, FCN16-VGG, FCN-Resnet50, FCN-ResNet101 respectively using FGSM, MI-FGSM, PGD. The maximum perturbation ϵ is set to 8 among all experiments, with pixel value in [0,255]. The number of iterations is 20 for MI-FGSM, and the decay factor μ is 1.0. For PGD attack, the iteration is 20, 100, 200. Note for FCN-ResNet50 and FCN-ResNet101, the iteration is 20 and 100 because it is very computationally expensive for PGD-200 used in complex model.

From Table1, we can observe that three different attack methods can efficiently drop the accuracy. For example, MI-FGSM drops the accuracy from 94.25% to 22.18% for FCN-ResNet101. PGD is the strongest among these three attack methods, and its accuracy decreases with the increment of number of iterations, which is supported by our analysis. In particular, an interesting phenomenon is that ResNet-based models not only achieve higher accuracy on clean input but also are more robust to adversarial examples.

Transfer Performance

In this section we show the transferability of adversarial examples generated from different models. In particular, we focus on MI-FGSM and PGD-20 attack method. Table 2 reports the accuracy of these models under attack. It is obvious that the transferability of MI-FGSM is better than PGD-20. In addition, adversarial examples are easier to transfer to models with the same backbone as the source model.

Number Of Iterations

Figure 3 illustrates the accuracy of FCN-ResNet50 under PGD attack with different number of iterations. We run $\{20, 40, 60, 80, 100\}$ iterations of PGD as our adversary,

Table 2: Transfer attack results for segmentation networks under MI-FGSM. The left column represents four basic segmentation models which are used to generate adversarial examples. Then examples generated on one model are transferred to attack others.* indicates the white-box attacks.

	Attack	FCN8-VGG	FCN16-VGG	FCN-Res50	FCN-Res101
FCN8-VGG	MI-FGSM	23.57*	25.43	**79.43**	**82.69**
	PGD-20	**17.20***	**20.69**	86.39	89.08
FCN16-VGG	MI-FGSM	24.70	22.13*	**79.47**	**82.83**
	PGD-20	**20.16**	**16.11***	86.61	89.09
FCN8-Res50	MI-FGSM	**82.80**	**82.49**	**16.24***	**69.20**
	PGD-20	88.96	88.76	16.46*	89.19
FCN8-Res101	MI-FGSM	**83.77**	**83.53**	**68.68**	22.18*
	PGD-20	89.08	88.89	87.64	**21.96***

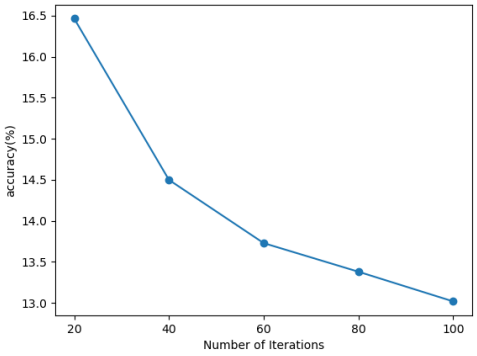

Figure 3: The accuracy of FCN-ResNet50 under PGD attack with different hyperparameters of number of iterations

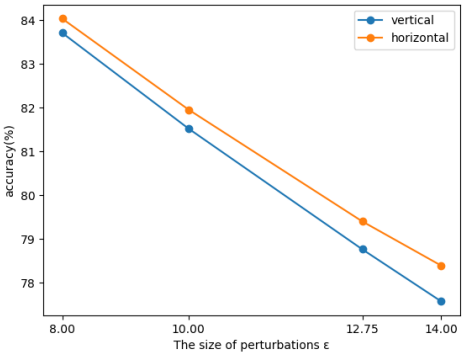

Figure 4: The accuracy of FCN-ResNet50 under square attack. The blue curve is square attack with vertical noise as initialization and the orange curve takes horizontal noise as initialization instead.

with step size $\alpha = \epsilon/iterations$. It can be seen that as the number of iterations grows, the accuracy of FCN-ResNet50 on VOC2012 test set drops. An intriguing phenomenon of the curve is that the scope of the accuracy curve decreases too. We highly assume the curve will be horizontal in the end.

Black-box Attack

For black-box attack, we take Square Attack into consideration which is the most efficient black-box attack to our knowledge. Figure 4 illustrates the results of horizontal and vertical initialization. It is not hard to find that the initialization of vertical is slightly better than horizontal initialization.

Conclusion

In this paper, we investigate the problem of generating adversarial examples for semantic segmentation leveraging the typical methods originally designed for classification. In particular, three gradient-based white-box attack and one query-based black-box attack are performed for segmentation models. Extensive experiment results verify that these methods work well on semantic segmentation, which might be attributed to the similarity between classification and segmentation.

While previous works have shown that adversarial attacks might be extended to the physical world and deceive face recognition systems, a practical attack against, e.g., an automated driving or surveillance system has not been presented yet. Investigating whether such practical attacks are feasible presents an important direction for future work. Furthermore, investigating whether other architectures for semantic segmentation are less vulnerable to adversarial perturbations is equally important.

References

Andriushchenko, M.; Croce, F.; Flammarion, N.; and Hein, M. 2020. Square attack: a query-efficient black-box adversarial attack via random search.

Arnab, A.; Miksik, O.; and Torr, P. H. 2018. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet a deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39(12):2481–2495.

Brendel, W.; Rauber, J.; and Bethge, M. 2018. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models.

Carlini, N., and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40(4):834–848.

Cheng, M.; Le, T.; Chen, P.-Y.; Yi, J.; Zhang, H.; and Hsieh, C.-J. 2018. Query-efficient hard-label black-box attack:an optimization-based approach.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88(2):303–338.

Fischer, V.; Kumar, M. C.; Metzen, J. H.; and Brox, T. 2017. Adversarial examples for semantic image segmentation.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition.

Hendrik Metzen, J.; Chaithanya Kumar, M.; Brox, T.; and Fischer, V. 2017. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Huang, Z., and Zhang, T. 2020. Black-box adversarial attack with transferable model-based embedding.

Ilyas, A.; Engstrom, L.; Athalye, A.; and Lin, J. 2018. Black-box adversarial attacks with limited queries and information.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ma, L.; Li, M.; Ma, X.; Cheng, L.; Du, P.; and Liu, Y. 2017. A review of supervised object-based land-cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing* 130:277–293.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2019. Towards deep learning models resistant to adversarial attacks.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, 506–519. New York, NY, USA: Association for Computing Machinery.

Paszke, A.; Chaurasia, A.; Kim, S.; and Culurciello, E. 2016. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc. 91–99.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Shen, G.; Mao, C.; Yang, J.; and Ray, B. 2019. Advspade: Realistic unrestricted attacks for semantic segmentation.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks.

Tao, A.; Sapra, K.; and Catanzaro, B. 2020. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*.

Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; and Yuille, A. 2017. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Zhang, Z.; Lin, H.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. 2020. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; and Torr, P. H. 2015. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, 1529–1537.