

# Semantically Stealthy Adversarial Attacks against Segmentation Models

Zhenhua Chen    Chuhua Wang    David Crandall  
Indiana University Bloomington  
Bloomington, IN 47405  
{chen478, cw234, djcran}@iu.edu

## Abstract

Segmentation models have been found to be vulnerable to targeted and non-targeted adversarial attacks. However, the resulting segmentation outputs are often so damaged that it is easy to spot an attack. In this paper, we propose semantically stealthy adversarial attacks which can manipulate targeted labels while preserving non-targeted labels at the same time. One challenge is making semantically meaningful manipulations across datasets and models. Another challenge is avoiding damaging non-targeted labels. To solve these challenges, we consider each input image as prior knowledge to generate perturbations. We also design a special regularizer to help extract features. To evaluate our model’s performance, we design three basic attack types, namely ‘vanishing into the context,’ ‘embedding fake labels,’ and ‘displacing target objects.’ Our experiments show that our stealthy adversarial model can attack segmentation models with a relatively high success rate on Cityscapes, Mapillary, and BDD100K. Our framework shows good empirical generalization across datasets and models.

## 1. Introduction

It is well-known that neural networks are vulnerable to adversarial attacks. For example, we can add imperceptible perturbations to an input image to manipulate the predictions of classifiers [8, 11, 26, 34, 36, 42], object detectors [40], segmentation models [1], object trackers [12, 18, 20, 38], edge detectors [7], 3D reconstruction models [14, 44], image caption models [41], face detectors [9], embodied agents [33], video classifiers [47], etc.

Among all the above adversarial attacks, those against segmentation models are relatively difficult because of the thousands of constraints that need to be considered for each input sample. So far, most adversarial segmentation models focus on non-targeted objectives [1, 40], which means that when a segmentation model is attacked, the abnormal predictions are noticeable. Although there are some targeted

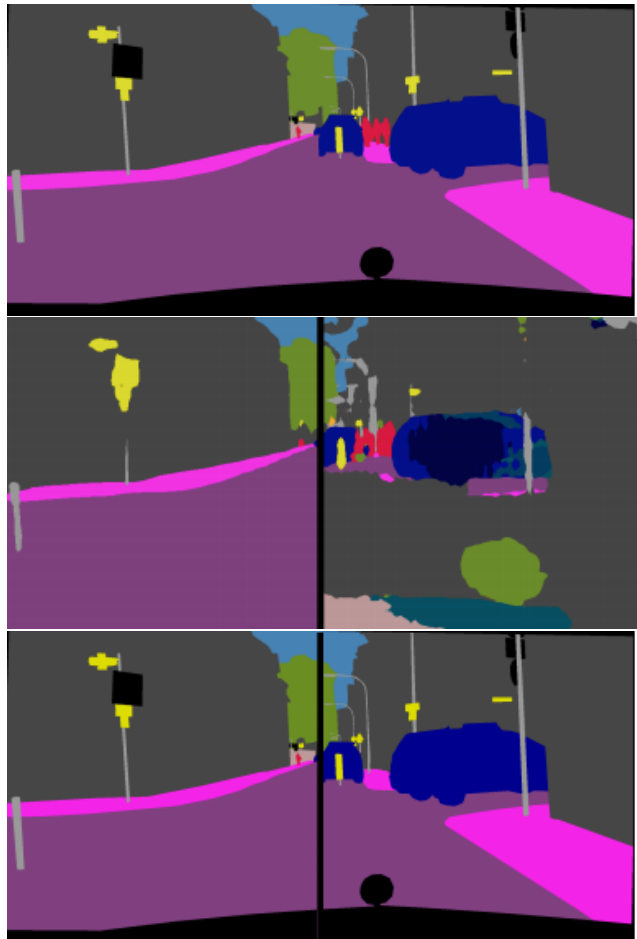


Figure 1. Comparison between stealthy and non-stealthy attacks. Top: Predictions without attacks. Red represents the ‘person’ label. Middle: Predictions before (left) and after (right) normal attacks in which the incorrect predictions are obvious. Bottom: Predictions before (left) and after (right) semantically stealthy attacks, in which it is difficult to detect an attack since the ‘person’ labels vanished into the environment.

adversarial segmentation models, most of them focus on either a static target [22, 29] or a particular attribute [22, 37].

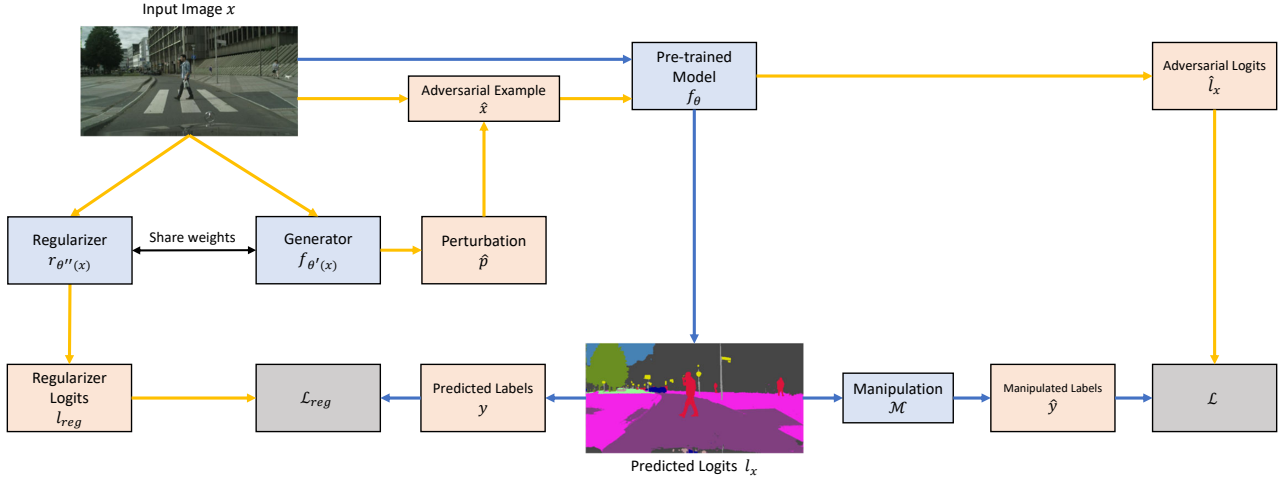


Figure 2. An overview of our attack model. During training, each input image is fed into both generator and regularizer to generate perturbation and regularizer logits. The scaled perturbation is then added to the original image to produce an adversarial example. Pre-trained target model receives both the original and adversarial image, and original predicted logits and adversarial logits are acquired. The predicted logits are manipulated based on different attack types, and output manipulated labels to calculate adversarial loss. The regularizer logits are paired with predicted labels to calculate regularizer loss.

Here we propose semantically “stealthy” adversarial attacks against segmentation models, so that they are not as noticeable. To illustrate this stealthy idea, we design three attack types, 1) making a target object vanish into the context; 2) adding a fake object; 3) replacing a target object with a fake one at a different position. For example, in Figure 4, the ‘person’ labels disappear after being attacked, and at the same time, these pixels are re-classified as ‘cars.’ Since the ‘car’ labels are common in this context, it would be difficult to detect this attack.

To achieve semantically stealthy adversarial attacks, we face two challenges. On one hand, the targeted labels need to be efficiently relabeled as designed. On the other hand, all other non-targeted pixels’ labels should be preserved. Since all the pixels’ labels in a segmentation model have to be considered, the total number of constraints is huge. To tackle this problem we introduce two approaches: 1) an image-dependent perturbation generation scheme, and 2) a regularizer. Our regularizer is a segmentation model using the same structure as FCN [21], as Figure 2 shows. The regularizer is introduced based on the following assumption: if a perturbation generator can generate effective noise then it should also be able to predict segmentation labels with high accuracy since the two tasks need common spatial features. In other words, the generated features should be task-irrelevant.

In addition to the task-related constraints, we also constrain our perturbations to be imperceptible by requiring their infinity norm to be smaller than 10,  $\|\hat{p}\|_{\infty} \leq 10$ ). We are also constrained by GPU memory: segmentation models

are usually large since they need to predict labels for every pixel, and we have to use another large neural network to fit the complex tasks. Thus, putting both networks during training would consume a large amount of memory. We introduce the concept of ‘attack efficiency,’ namely the ratio of the attack/target model’s size.

Our primary contributions are summarized as follows:

- We propose an attack framework that can create stealthy, semantically-meaningful attacks against segmentation models.
- We design three basic types of semantic attacks: removing labels, adding fake labels, and moving labels. Other high-level semantic manipulations can be performed by combining these basic types.
- We evaluate our model’s generalization ability across different datasets.
- We propose the concept of parameter-wise efficiency for adversarial models.

## 2. Related work

### 2.1. Universal adversarial attacks

The universal adversary might be a solution to overcome the computationally intensive issue of FGSM-based attacks. Universal Adversarial Perturbations (UAP) were introduced by [23], and can generate sample-agnostic perturbations of-fine and thus can apply attacks in real-time. The key idea of UAP is to find a universal perturbation that can fool all

the training/test images. UAP was originally designed for classifiers but can be easily extended to attack segmentation models by, for example, finding a perturbation that maximizes the predictive probability of a target class by iteratively backpropagating the gradients to the input space [22]. When the training is finished, the final gradients form the perturbation which is added to each input image during testing time. Another way to obtain UAP is by calculating singular vectors of the Jacobian matrices of the feature maps [27]. It is also possible to use a neural network to generate UAP [15, 24, 29]. Benz et al. [2], Gupta et al. [13], Zhang et al. [45] explore the idea of generating UAPs that only attack a group of classes and limit the influence on the remaining classes. Another interesting perspective towards UAP is considering images rather than perturbations as noise [3]. As for defending against UAP, adversarial learning is still the most efficient way [46].

## 2.2. Image-dependent adversarial attacks

Methods based on UAP attack models largely avoid the computationally expensive issue. However, they are not so flexible since each input image, though it can be considered as prior knowledge, is ignored. One straightforward way to solve this issue is by generating perturbations from input images. We can also adopt a GAN-like structure, in which the discriminator corresponds to the target model while the generator is supposed to generate the perturbations. The perturbation generator can start from either random noise or an image. If the perturbation comes from an image, it is called an image-dependent perturbation. For example, Poursaeed et al. [29] proposes a universal, GAN-like attack model that can attack both segmentation models and classifiers.

## 2.3. Semantic adversarial attacks

Semantic/targeted adversarial attacks are related to our work. Shamsabadi et al. [32] propose a content-based black-box adversarial attack through manipulating background colors. Joshi et al. [19] creates semantic adversarial attacks against classifiers by manipulating specific attributes. These types of attacks are somewhat stealthy but their target models are classifiers. Treu et al. [37] proposed a method to overlay generated adversarial texture on the clothing of a person to fool segmentation networks. Metzen et al. [22] is most alike our work in which an adversarial attack for hiding person labels is proposed. However, Metzen et al. [22] adopt universal perturbations and tested only on one dataset.

## 2.4. Physical adversarial attacks

Apart from adding perturbations, it is also possible to add physical attributes to a scene in order to attack. For example, Sharif et al. [35] can fool a face recognizer by

adding glasses to a face, while Qiu et al. [30] manipulates face attributes (like hair color, etc.). Duan et al. [10] manipulate classifiers’ predictions by camouflaging physical adversarial examples into natural styles. These methods used to generate physical adversarial examples are weakly related to our work.

## 3. Our Approach

Many studies [5, 39, 40] show that complex tasks such as semantic segmentation are vulnerable to adversarial attacks, and that generated subtle perturbations to inputs can completely break the prediction outputs. We present a stealthy adversarial attack approach that not only can hide an attack behavior, but also keeps other non-targeted labels correctly classified. We realize our goal by designing a regularized pipeline to generate image-dependent perturbations, as shown in Figure 2.

In this section, we first give an overview of the adversarial attack problem and our proposed stealthy approach. We then introduce how the generator and regularizer work.

### 3.1. Problem definition

Let  $f_\theta$  be the target segmentation model trained on some dataset with (image, label) pairs  $\mathcal{D} = \{(x, y)\}$  where  $x \in \mathbb{R}^{h \times w \times c}$ ,  $y \in \mathbb{R}^{h \times w}$  and  $h$ ,  $w$ , and  $c$  represent the height, width, and number of image channels, respectively. The perturbation  $\hat{p} \in \mathbb{R}^{h \times w \times c}$  has the same size as the input image that can be acquired by the generator  $f_{\theta'}(x)$ . An adversarial example  $\hat{x} \in \mathbb{R}^{h \times w \times c}$  is the addition of the perturbation  $\hat{p}$  and the input image  $x$ . We need to map both  $x$  and  $\hat{x}$  to logits through  $f_\theta$  during training,

$$l_x = f_\theta(x), \hat{l}_x = f_\theta(\hat{x}). \quad (1)$$

Apart from  $l_x$  and  $\hat{l}_x$ , we also train a regularizer that is supposed to predict segmentation labels,

$$l_{reg} = r_{\theta''}(x). \quad (2)$$

Finally, we pair  $l_x, \hat{l}_x, l_{reg}$  with their corresponding labels  $y, \hat{y}, y$  to calculate the loss.

### 3.2. Model overview

As Figure 2 shows, our framework contains a target model, a generator, and a regularizer. Input images  $x$  are fed into the generator  $f_{\theta'}$  to produce perturbations. The regularizer (which shares the same backbone  $f_{\theta'}$  with the generator) is trained to learn the segmentation task. Both original input images and the corresponding perturbed images are paired and fed to the target model  $f_\theta$  to generate logits  $l_x, \hat{l}_x$ . After we map  $\hat{l}_x$  to  $\hat{y}$ ,  $l_x$  and  $\hat{y}$  are fed into the loss function for training.

During training, we first load the pre-trained target model  $f_\theta$  and freeze it to prevent the weights from updating during back-propagation. Then the generator takes an image sample and outputs a perturbation image. The regularizer generates logits to fit the original segmentation task. The adversarial image is obtained by adding the scaled perturbation to the original image. Both the original and adversarial are fed into  $f_\theta$  and output the original and the adversarial logits respectively. Depending on attack designs, we pair manipulated labels with the predicted labels to calculate the adversarial loss. At the same time, we pair the logits generated by the regularizer with the ground truth segmentation labels to calculate the regularizer loss. Finally, both cross-entropy losses from the regularizer and the generator are backpropagated to the generator space. The whole algorithm is summarized in Algorithm 1.

### 3.2.1 Generator & regularizer

We adopt two versions of generators to deal with different sizes of segmentation models: a Unet [31] network (as shown in Figure 13), and a FCN network structure [21] with ResNet-101 backbone [16]. The generator  $f_{\theta'}$  is to map an input image to perturbations, as shown in Figure 13. At the same time, we have a regularizer  $r_{\theta''}$  which is responsible for predicting the semantic segmentation labels of the original image. The potential assumption is that if a generator can generate perturbations that serve a particular target, then the embedded features that come from the generator should also have a good knowledge of the spatial relationship within each input image. In other words, the two branches of the generator (one for semantic segmentation label, the other one for perturbations) are complementary with each other. After the perturbation is acquired, we forward both the original image and the adversarial image through the target model to obtain predicted and adversarial labels. We use a cross-entropy loss to regulate the generator and help to preserve labels of non-targeted pixels. The adversarial labels are manipulated (see Section 3.2.2) and used to calculate the adversarial loss (see 3.2.3 for details).

### 3.2.2 Attack types

In general, we achieve stealthiness via manipulating targeted pixels' labels as designed and at the same time preserving untargeted pixels' labels. During attack, we map 'clean' labels to 'semantically stealthy' labels,

$$\hat{y} = \mathcal{M}(l_x) \quad (3)$$

We design three attack types to manipulate the predicted labels and validate our idea. We believe that these three attack types are basic and combining these attack types can be used to generate other higher-level semantically attack

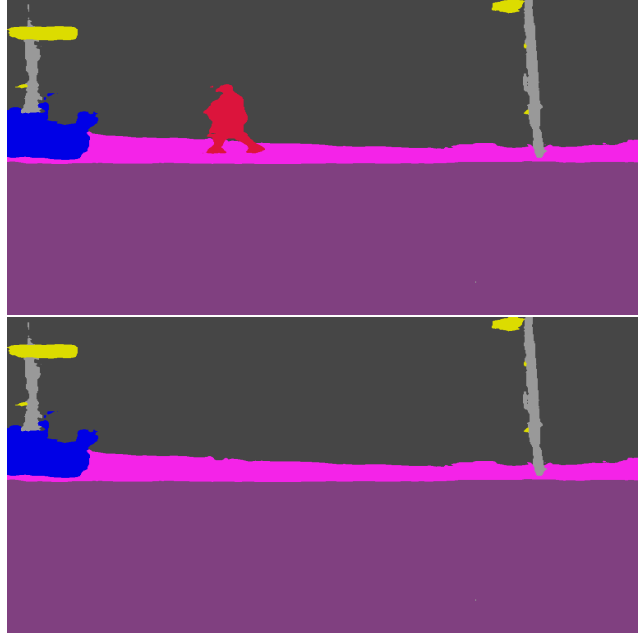


Figure 3. How we convert the original predictions to the ‘adversarial ground truth.’ The ‘person’ labels (the red area) near the wall are converted to label ‘wall’ while those near the sidewalk are converted to label ‘sidewalk.’

types: Type #1, remove the target class, Type #2, generate a class label from an external source, and Type #3, combine the previous two types. More specifically,

**Type #1:** The target model is made to fail to predict a designated class. As a result, the original label has disappeared or blended into the surrounding context. The most straightforward way of achieving this goal is to assign each target pixel its neighbor’s label by clustering. However, clustering is very time-consuming. We thus adopt a heuristic approach by replacing each target pixel’s original label with its second likely label in the original prediction. Practically, it works well, as Figure 3 shows.

**Type #2:** The labels of a set of referred pixels are transformed into a specific class. Visually, this makes it appear like an attacker creates a set of new labels that comes from nowhere. We manipulate the predicted label by adding a new mask from an external source.

**Type #3:** This task can be achieved by combining the attacks of Type #1 and Type #2: removing the target class labels in one position and then create fakes class labels in another position.

### 3.2.3 Loss

We design our loss function to make sure that the target label is manipulated from  $y$  to  $\hat{y}$ , as Equation 4 shows.  $T_k$

---

**Algorithm 1:** Semantically Stealthy Adversarial Attack.

---

**Input :** input image  $x$ , Target model  $f_\theta$  with parameters  $\theta$ . Generator  $f_{\theta'}$  with parameters  $\theta'$ , regularizer  $r_{\theta''}$  with parameters  $\theta''$ , Stealthy label mapper  $\mathcal{M}$ .

**Output :** Adversarial perturbations  $\hat{p}$ , Adversarial input image  $\hat{x}$

Initialize the generator  $f_{\theta'}$ , Initialize the regularizer  $r_{\theta''}$

**while not converge do**

    Generate regularizer logits  $l_{reg} = r_{\theta''}(x)$ , and perturbations  $\hat{p} = f_{\theta'}(x)$ ;

    Obtain an adversarial image  $\hat{x} = x + \hat{p}$ ;

    Generate the clean logits of the target model  $l_x = f_\theta(x)$ ;

    Generate the dirty logits of target model  $\hat{l}_x = f_\theta(\hat{x})$ ;

    Create semantically stealthy labels  $\hat{y} = \mathcal{M}(l_x)$ ;

    Calculate the regularizer loss  $\mathcal{L}_{reg}(y, r_{\theta''}(x))$ ;

    Calculate the adversarial loss  $\mathcal{L}(l_x, \hat{l}_x)$ ;

    Back-propagate the gradients;

---

equals 1 if the pixel  $k$  is one of our target pixels, and otherwise is 0. Variables  $n, c', m$  are the batch size, channel size (total number of categories), and the total number of pixels in each image, while  $p_{i,j,k}$  represents the probability of each pixel belonging to  $j$  which comes from either  $l_x \in \mathbb{R}^{m \times c'}$  (the logits forms of  $y$ ) or  $\hat{l}_x \in \mathbb{R}^{m \times c'}$ . The loss function  $\mathcal{L}(\cdot)$  is shown in Equation 4.

$$\mathcal{L}(\hat{l}_x, \hat{y}) = \min \sum_{i=0}^n \sum_{k=0}^m T_k(-\log p_{i, \hat{y}_k, k}) + (1 - T_k)(-\log(p_{i, \hat{y}_k, k})). \quad (4)$$

Assuming the logits that come from the generator is  $l_{reg}$ , then the regularizer loss can be summarized as,

$$\mathcal{L}_{reg}(l_{reg}, y) = \min \sum_{i=0}^n \sum_{k=0}^m -\log p_{i, y_k, k}. \quad (5)$$

We introduce  $\lambda_0$  to be the weight of regularizer loss. The total loss term is,

$$\mathcal{L}_{total} = \mathcal{L} + \lambda_0 * \mathcal{L}_{reg}. \quad (6)$$

It is common practice to limit the infinity norm of the perturbations to achieve imperceptibility, as shown in [11]. If not specified,  $\xi$  is always equal to 10,

$$\|\hat{p}\|_\infty \leq \xi. \quad (7)$$

## 4. Experiments

### 4.1. Datasets & Metrics

We choose three street-view datasets as the testing set since all our attack types focus on pedestrians. Specifically,

we use Cityscapes (2975 training images, 1525 testing images) [6], Mapillary Vistas Dataset (18K training images, 5K testing images) [25], and BDD100K (7K training image, 2K testing images) [43].

We evaluate our model by using *success rate* as the performance metric. Each success rate is divided into two parts: manipulated and preserved rate. Manipulated rate is defined as the percentage of labels that are manipulated successfully for targeted pixels while the preserved rate is defined as the percentage of labels that are preserved for non-targeted pixels.

### 4.2. Implementation details

We implement our network in PyTorch [28] and perform all experiments with Nvidia Titan Xp Pascal GPUs. In Section 4.5, we use FCN-8s as the target segmentation model, and pre-train the model on all three datasets. We resize each input image to  $512 \times 1024$  during training. For Section 4.6, we use CityScapes as our train/test dataset. Each input image is loaded at its original size and then cropped randomly to  $768 \times 768$  during training. For all the models, we choose *Person* and *Rider* as our target labels. The weight of our regularizer  $\lambda_0$  is set to 1e-2. The learning rate is 1e-4, and batch size is set to 16.

### 4.3. Quantitative analysis

Table 1 shows the manipulated/preserved rate of attacking the target model [21]. We observe that manipulating non-targeted labels into target labels (Type #2) seems to be easier than blending target labels into non-targeted labels (Type #1). We conjecture that the model does not need to consider the context information in the former task, thus leading to a higher manipulated rate. However, our results show that preserving non-targeted labels in attack type #2 is more difficult than in attack type #1. We speculate that this

Dataset	Type#1	Type#2	Type#3
Cityscapes	74.71% / 91.40%	91.27% / 92.76%	67.73% / 88.74%
BDD100K	66.56% / 97.09%	68.69% / 92.46%	66.54% / 90.58%
Mapillary	68.39% / 96.41%	78.67% / 88.62%	56.07% / 85.42%

Table 1. The performance of our framework on three street-view datasets. Each item represents success rates for type ‘manipulated’ / ‘preserved’ in one of the three datasets. For example, the top-left item (red-colored) means that the attack model trained on Cityscapes can make 74.71% targeted pixels vanish into the background while making 91.40% non-targeted pixels preserve their original labels. All the target models are trained by FCN.

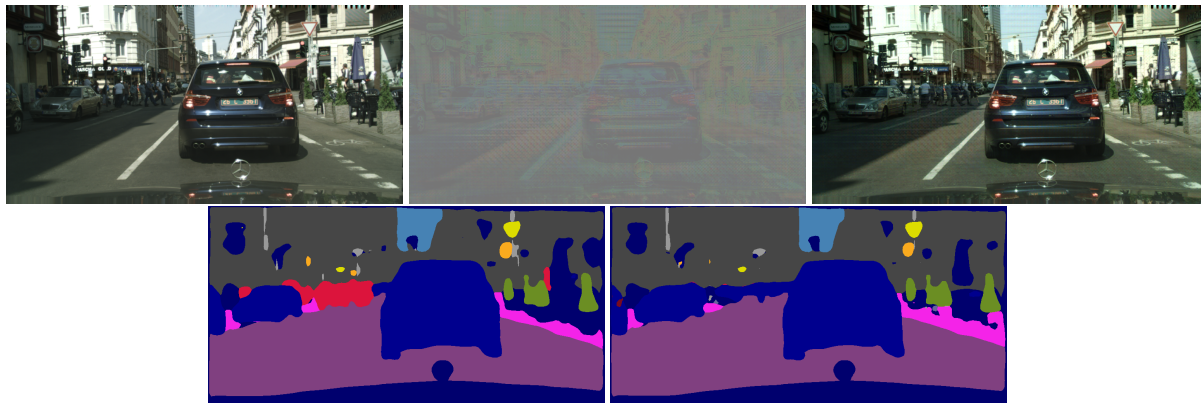


Figure 4. Type #1 attack on Cityscapes. The ‘person’ labels vanish into the background. Top left: The input image. Top middle: The perturbations. Top right: The input image + perturbations. Lower left: Normal predictions. Lower right: Predictions after attacks.

might be because it is challenging to preserve surrounding information when the manipulated pixels do not fit in context. The results also suggest that combining the two tasks (type #3) decreases the overall performance.

#### 4.4. Qualitative analysis

Figure 4 shows the results of Type #1 attack. The person labels are converted to building, car, sidewalk, and road labels. There are some ‘residuals’ left in the original position. Figure 5 shows the results of Type #2 attack. We use the binary mask of person label to train our generator. The perturbation includes highlights around the boundary of added person labels, and the target model classifies these non-targeted pixels as person labels. Figure 6 shows the results of combining the previous two types of attacks. The person labels in the original prediction are not completely removed and three fake ‘person’ regions are embedded. Comparing to the above two tasks, Type #3 is the most difficult among them.

#### 4.5. Generalization across datasets

It is desirable to check the generalization ability of our model across different datasets. In other words, we want to find out whether our model can achieve good performance on unseen distributions. Specifically, we train an adversarial model (model #1 in Table 5) on one dataset (e.g. Cityscapes) to attack the segmentation models trained on

Datasets	Cityscapes	BDD100K	Mapillary
Cityscapes	96.34%	95.44%	92.96%
BDD100K	97.24%	97.86%	96.56%
Mapillary	96.86%	96.09%	97.41%

Table 2. Our model’s generalization across datasets on Type#1 attack. Each item is the success rate within/across datasets. For example, the red-colored 95.44% indicates that the attack model trained on Cityscapes can achieve a success rate of 95.44% against a segmentation model trained on BDD100K.

two other datasets (e.g. Mapillary, and BDD100K).

We can draw two conclusions here from Table 2: each attack model performs best within its distribution, and each attack model’s performance generalizes well on unseen datasets (distributions). For example, the results of BDD100K and Mapillary show that both models trained on these two datasets are able to reach comparable performance when generalizing to others. We conjecture that a larger dataset may contain more useful information during training, thus resulting in better generalization ability.

#### 4.6. Generalization across models

Generalization across models is also an important aspect of an adversarial model. We select two target models based on DeepLabV3Plus [4]. One has a backbone of MobileNets [17], while the other one has a backbone of

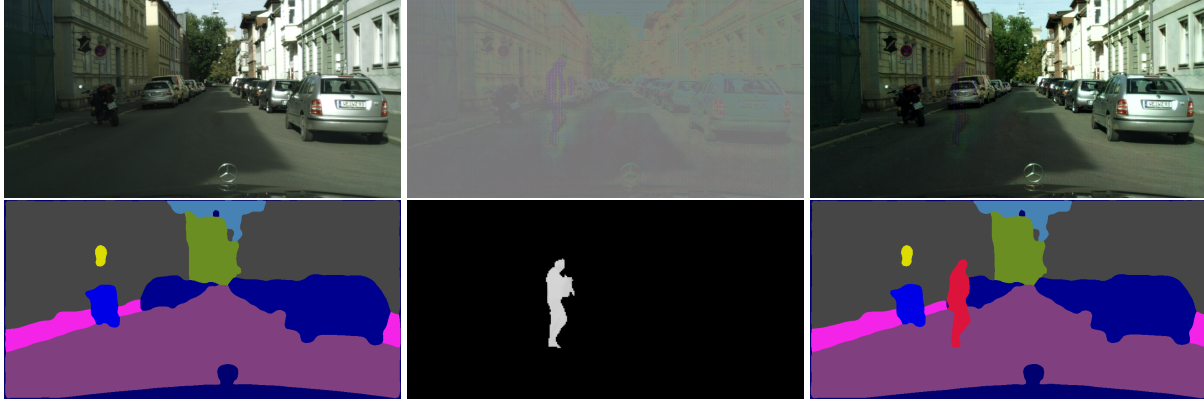


Figure 5. Type #2 attack on Cityscapes. Fake ‘person’ labels are embedded into the prediction domain. Top left: The input image. Top middle: The perturbations. Top right: The input image + perturbations. Lower left: Normal predictions. Lower middle: The fake label mask. Lower right: Predictions after attacks.



Figure 6. Type #3 attack on Cityscapes. The target ‘person’ labels disappear and the fake ‘person’ labels are embedded. Top left: The input image. Top middle: The perturbations. Top right: The input image + perturbations. Lower left: Normal predictions. Lower middle: The fake label mask. Lower right: Predictions after attacks.

Models	DLV3+-MobileNets	DLV3+-ResNet101
Type #1		
DLV3+-MobileNets	84.36%	<b>89.16%</b>
DLV3+-ResNet101	85.49%	89.92%
Type #2		
DLV3+-MobileNets	87.42%	90.67%
DLV3+-ResNet101	96.63%	96.80%
Type #3		
DLV3+-MobileNets	80.91%	83.35%
DLV3+-ResNet101	82.82%	84.80%

Table 3. Our model’s generalization across attack models trained on CityScapes. Each item is the overall success within/across models. For example, for type #1 attack, the red-colored **89.16%** indicates that the attack model trained on DeepLabV3Plus-MobileNets can achieve a success rate of **89.16%** against a target model pretrained on DeepLabV3Plus-ResNet101.

ResNet. Since the target model is very deep, we replace the

generator with a deeper version, a ResNet-based FCN. As Table 3 shows, for all three attack types, the success rates are quite close when the two target models are switched.

#### 4.7. Ablation study

The regularizer in our model plays a vital role in attacking segmentation models dynamically. The potential assumption is that if a generator can create perturbations for dynamic target labels, then the embedded features that come from the generator should also have a good knowledge of the spatial structure within each image. As a result, we can use these embedded features to generate labels that come from the target model. To explore the role of the regularizer, we compare the performance (manipulated rate) between our models and the corresponding regularizer-free ones on three datasets. As Table 4 shows, the regularizer can improve the manipulated rate by around 12% in Cityscapes and Mapillary, and 8% in Mapillary.

Model type	Cityscapes	BDD100K	Mapillary
Normal	74.71%	66.56%	68.39%
Regularizer-removed	62.87%	54.88%	59.98%

Table 4. The success rate of our model (as shown in Figure 13) with/without the proposed regularizer across different datasets.

attack model size	target model size	ratio	Cityscapes	BDD100K	Mapillary
Model#1 (531M)	269M	1.974	96.34%	97.86%	96.09%
Model#2 (10.69M)	269M	0.040	74.71%	66.56%	68.39%
Model#3 (2.69M)	269M	0.010	64.47%	53.63%	60.39%
Model#4 (692K)	269M	0.003	60.63%	50.99%	45.53%

Table 5. The parameter-wise efficiency of our model on type #1 attack. Model #1 adopt the generator in [29]. Model #2 uses the generator in Figure 13. Model #3 is the same as Model #2 only the number of feature maps in each layer is cut in half. Similarly, Model #4 cuts half of its parameters from Model #3.

#### 4.8. Efficiency of attack models

The efficiency of attack models has been largely ignored by previous works. However, as attack/target models become larger and more complex, the efficiency issue can no longer be ignored. For example, the success rate is higher when using a ResNet-based generator to attack a MobileNets-based target model compared to vice versa.

Here we propose a metric that is defined as the ratio of attack models’ sizes and the target models’ sizes. For the same manipulated rate, smaller attack models are more efficient. We list four different models and show the total number of parameters of them in Table 5 in descending order.

We can see that larger attack models can achieve a higher success rate. For example, when the model is twice as large as the target model, model #1 can achieve an almost perfect manipulated rate (more than 96%). With only 4% parameters, our model #2 can achieve around 70% manipulated rate. It is interesting that with less than 0.3% parameters (model #4), our model can still achieve relatively good performance.

### 5. Conclusion

We propose a framework for creating semantically stealthy adversarial attacks against segmentation models. In particular, we design an algorithm that can manipulate targeted pixels’ labels as designed while keeping other pixels’ labels untouched. To achieve this goal, we introduce prior knowledge (each perturbation is conditionally dependent on the corresponding input image) as well as a special regularizer into our attack model. We evaluate our model’s performance by success rates on three types of stealthy attacks. The experiment shows that our framework has a relatively high success rate across datasets/models. Finally, we propose the concept of attack efficiency which may help estimate attack models’ parameter-wise efficiency.

### References

- [1] Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. On the robustness of semantic segmentation models to adversarial attacks. *CoRR*, abs/1711.09856, 2017. URL <http://arxiv.org/abs/1711.09856>.
- [2] Philipp Benz, Chaoning Zhang, Tooba Imtiaz, and In So Kweon. Double targeted universal adversarial perturbations. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.
- [3] Tooba Imtiaz Chaoning Zhang, Philipp Benz. Understanding adversarial examples from the mutual influence of images and perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR*, abs/1802.02611, 2018. URL <http://arxiv.org/abs/1802.02611>.
- [5] Moustapha M Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/d494020ff8ec181ef98ed97ac3f25453-Paper.pdf>.
- [6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.



- [7] Christian Cosgrove and Alan Yuille. Adversarial examples for edge detection: They exist, and they transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.
- [8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [9] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [10] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A. K. Qin, and Yun Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *CVPR*, 2020.
- [11] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. *ArXiv e-prints*, December 2014.
- [12] Qing Guo, Xiaofei Xie, Felix Juefei-Xu, Lei Ma, Zhongguo Li, Wanli Xue, Wei Feng, and Yang Liu. Spark: Spatial-aware online incremental attack against visual tracking. In *ECCV*, 2020.
- [13] Tejus Gupta, Abhishek Sinha, Nupur Kumari, Mayank Singh, and Balaji Krishnamurthy. A method for computing class-wise universal adversarial perturbations, 2019.
- [14] Abdullah Hamdi, Sara Rojas, Ali Thabet, and Bernard Ghanem. Advpc: Transferable adversarial perturbations on 3d point clouds. In *ECCV*, 2020.
- [15] Jamie Hayes and George Danezis. Learning universal adversarial perturbations with generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 43–49, 2018. doi: 10.1109/SPW.2018.00015.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
- [17] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. URL <http://arxiv.org/abs/1704.04861>.
- [18] Shuai Jia, Yibing Song, Chao Ma, and Xiaokang Yang. Iou attack: Towards temporally coherent black-box adversarial attack for visual object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [19] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: parametric transformations that fool deep classifiers. In *ICCV*, 2019.
- [20] Siyuan Liang, Xingxing Wei, Siyuan Yao, and Xiaochun Cao. Efficient adversarial attacks for visual object tracking. In *ECCV*, 2020.
- [21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014. URL <http://arxiv.org/abs/1411.4038>.
- [22] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2774–2783, Oct 2017. doi: 10.1109/ICCV.2017.300.
- [23] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. *CoRR*, abs/1610.08401, 2016. URL <http://arxiv.org/abs/1610.08401>.
- [24] Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R. Venkatesh Babu. Nag: Network for adversary generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 742–751, 2018. doi: 10.1109/CVPR.2018.00084.
- [25] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *International Conference on Computer Vision (ICCV)*, 2017. URL <https://www.mapillary.com/dataset/vistas>.
- [26] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897, 2014. URL <http://arxiv.org/abs/1412.1897>.
- [27] Ivan Oseledets and Valentin Khruikov. Art of singular vectors and universal adversarial perturbations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8562–8570, 2018. doi: 10.1109/CVPR.2018.00893.
- [28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [29] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

- [30] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditional image editing, 2020.
- [31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.
- [32] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *CVPR*, 2020.
- [33] shan Liu, Tairan Huang, Xianglong Liu, Yitao Xu, Yuqing Ma, Xinyun Chen, Stephen J. Maybank, and Dacheng Tao. Spatiotemporal attacks for embodied agents. In *ECCV*, 2020.
- [34] Mahmood Sharif, Sruti Bhagavatula, Lujio Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS '16*, pages 1528–1540, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4139-4. doi: 10.1145/2976749.2978392. URL <http://doi.acm.org/10.1145/2976749.2978392>.
- [35] Mahmood Sharif, Sruti Bhagavatula, Lujio Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS '16*, pages 1528–1540, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4139-4. doi: 10.1145/2976749.2978392. URL <http://doi.acm.org/10.1145/2976749.2978392>.
- [36] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013. URL <http://arxiv.org/abs/1312.6199>.
- [37] Marc Treu, Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Fashion-guided adversarial attack on person segmentation, 2021.
- [38] Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, and Liang Lin. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [39] Chaowei Xiao, Ruizhi Deng, Bo Li, Fisher Yu, Mingyan Liu, and Dawn Song. Characterizing adversarial examples based on spatial consistency information for semantic segmentation, 2018.
- [40] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. Adversarial examples for semantic segmentation and object detection. *CoRR*, abs/1703.08603, 2017. URL <http://arxiv.org/abs/1703.08603>.
- [41] Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. Exact adversarial attack to image captioning via structured output learning with latent variables. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [42] Zhewei Yao, Amir Gholami, Peng Xu, Kurt Keutzer, and Michael W. Mahoney. Trust region based adversarial attack on neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [43] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *CoRR*, abs/1805.04687, 2018. URL <http://arxiv.org/abs/1805.04687>.
- [44] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L. Yuille. Adversarial attacks beyond the image space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [45] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon. Cd-uap: Class discriminative universal adversarial perturbation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6754–6761, Apr. 2020. doi: 10.1609/aaai.v34i04.6154. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6154>.
- [46] Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. A survey on universal adversarial attack. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4687–4694. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/635. URL <https://doi.org/10.24963/ijcai.2021/635>. Survey Track.
- [47] Hu Zhang, Linchao Zhu, Yi Zhu, and Yi Yang. Motion-excited sampler: Video adversarial attack with sparkled prior. In *ECCV*, 2020.