NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

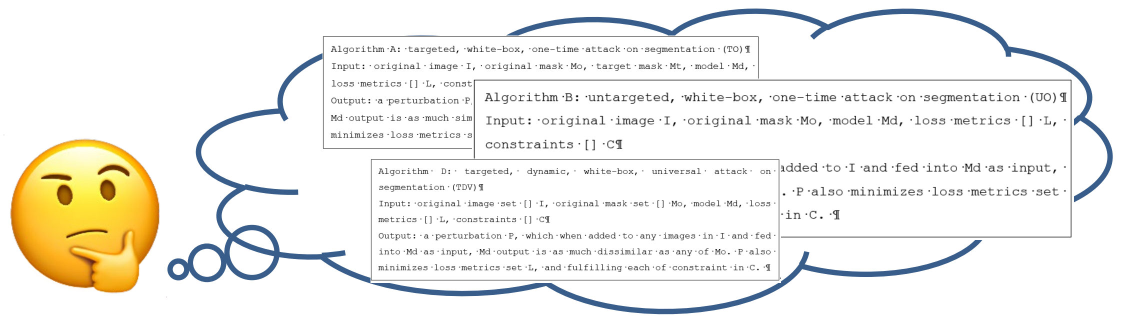
SCSE22-0052

A segmentation adversarial attack pipeline and algorithm implementation

Student: Zhao Peizhu Supervisor: Prof Tan Rui & Dr. Guo Dongfang

What are the problems?

- Adversarial attacks are serious threat which needs intense study
- Adversarial attacks are hard problems for Autonomous Vehicle (AV) development
- Attacks on segmentation task are proved can be devastating to computer vision systems
- Lack of integrated pipelines for segmentation attack researches



```
Algorithm A: targeted, white-box, one-time attack on segmentation (TO)
Input: original image I, original mask Mo, target mask Mt, model Md,
loss metrics [] L, const
Output: a perturbation P,
Md output is as much sim
minimizes loss metrics s

Algorithm B: untargeted, white-box, one-time attack on segmentation (UO)
Input: original image I, original mask Mo, model Md, loss metrics [] L,
constraints [] C
added to I and fed into Md as input,
. P also minimizes loss metrics set
in C.

Algorithm D: targeted, dynamic, white-box, universal attack on
segmentation (TDV)
Input: original image set [] I, original mask set [] Mo, model Md, loss
metrics [] L, constraints [] C
Output: a perturbation P, which when added to any images in I and fed
into Md as input, Md output is as much dissimilar as any of Mo. P also
minimizes loss metrics set L, and fulfilling each of constraint in C.
```

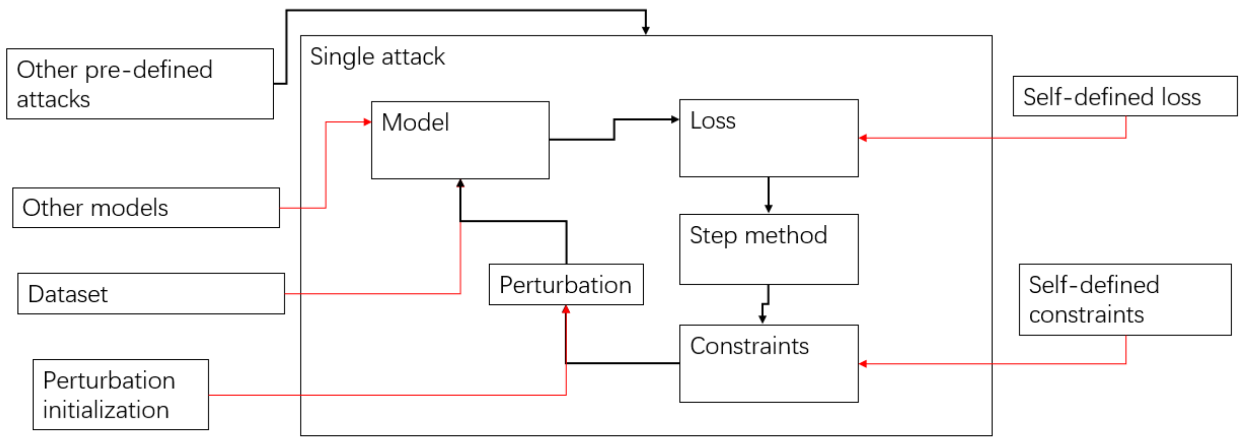
What are our objectives?:

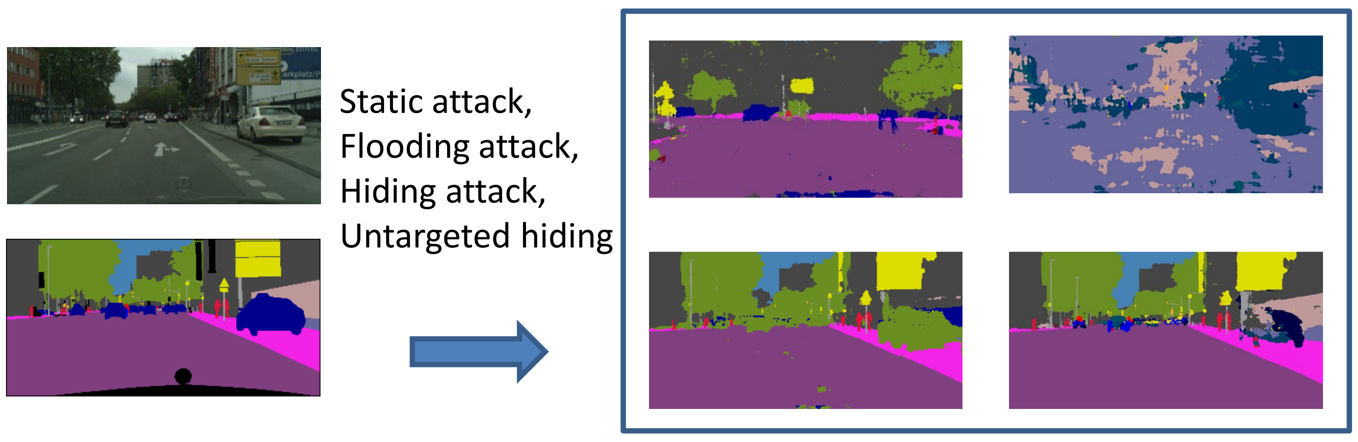
- Construct a modularized attack pipeline for segmentation adversarial attack tasks
- Implement the algorithm for most common attacks and evaluate their effects

What can it do?

It can attack computer visual perception in various ways:

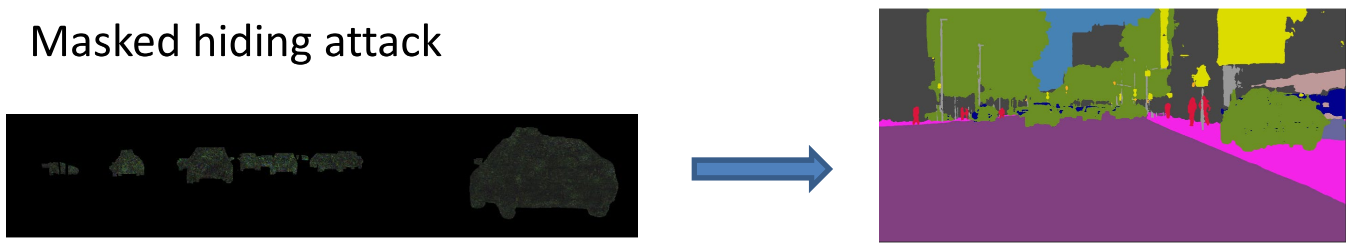
- Masking cars
- Disable it by flooding it with garbage
- Quickly destroy some important objects (e.g. cars)
- Fool a static camera with another static scene
- Apply a perturbation to any input that fools the camera with a static scene
- Apply a perturbation to any input that hiding some elements (e.g. cars) from that input
- … And other more





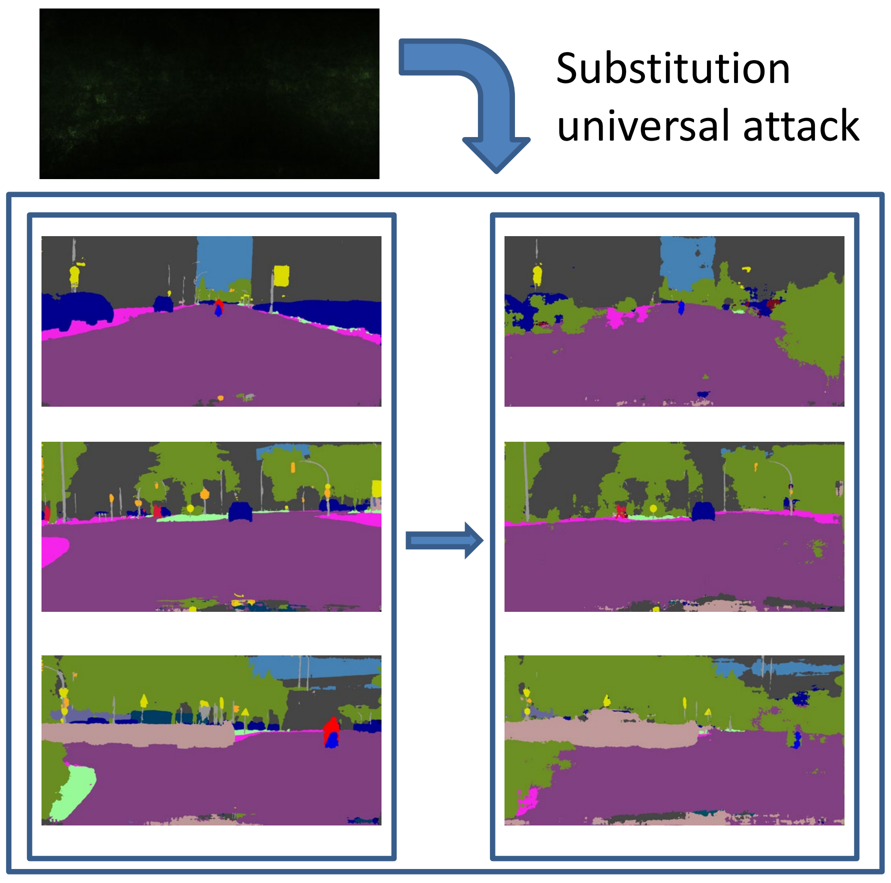
Static attack,
Flooding attack,
Hiding attack,
Untargeted hiding

Masked hiding attack

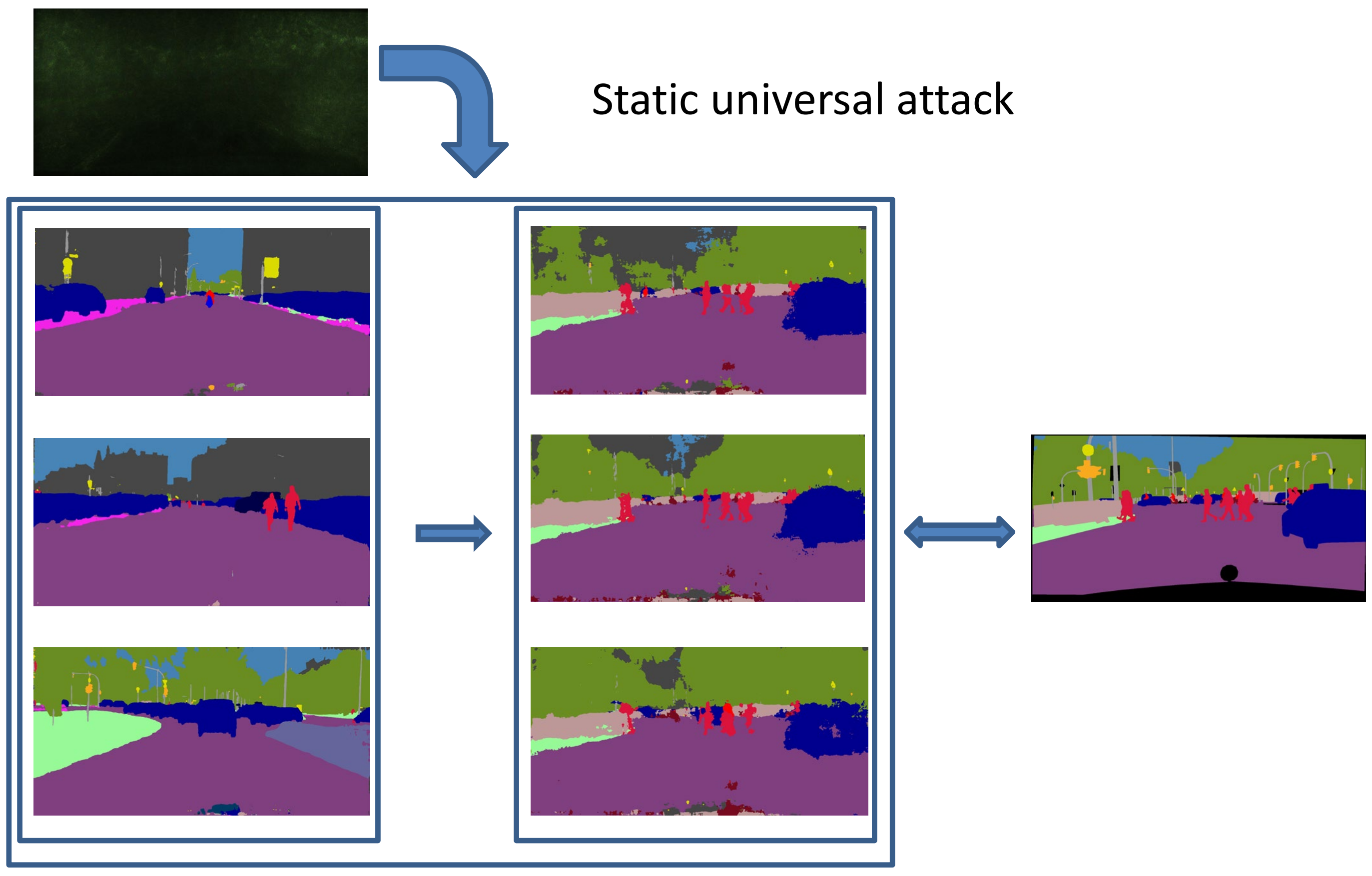


Why is it important?

- White box attacks are testbeds for almost all attacks
- Physically resizable attacks are mostly based on transferred white boxes!
- Such attacks can be applied to not only AV scenarios but also IoT systems
- A proved stealthy and scalable security threat



Substitution universal attack



Static universal attack