Backdoor attacks against semantic segmentation models

Author: Dominik Jambrović

Mentor: prof. dr. sc. Siniša Šegvić

Table of contents

Introduction

AI security:

- a lot of research concerning “classic” classification models
- significantly less research concerning **semantic segmentation models**

Potential threats:

- adversarial attacks
- **backdoor attacks**

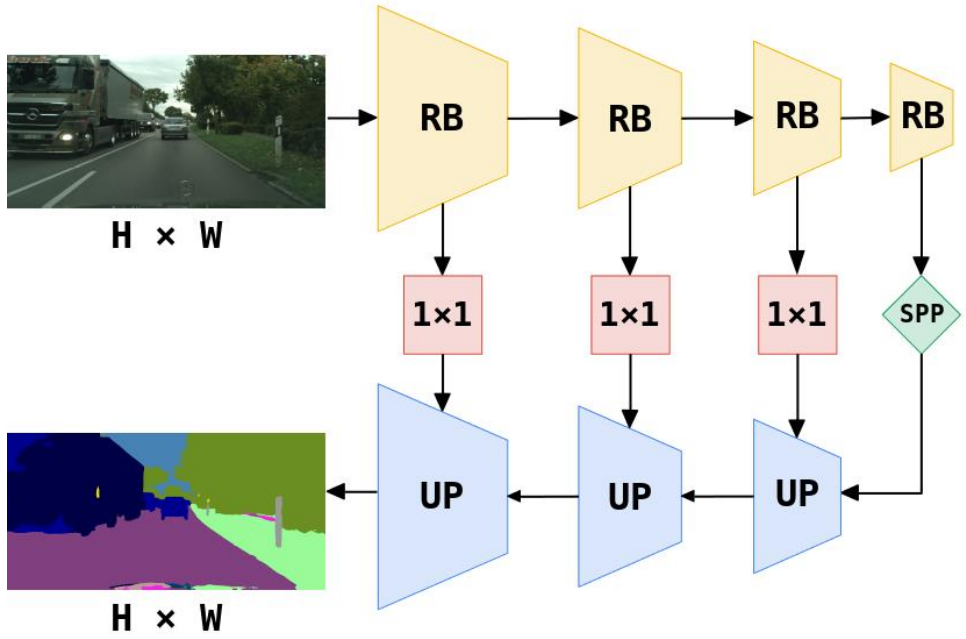
Backdoor attacks

- images with an added **trigger** coupled with **altered labels**
- goal: embedding a **backdoor** in the targeted model

Experiments – Hidden backdoor attack

Experimental setup

- dataset: **ADE20k**
- architecture: **Single-Scale SwiftNet**
- number of epochs: **200**



Attack types - triggers

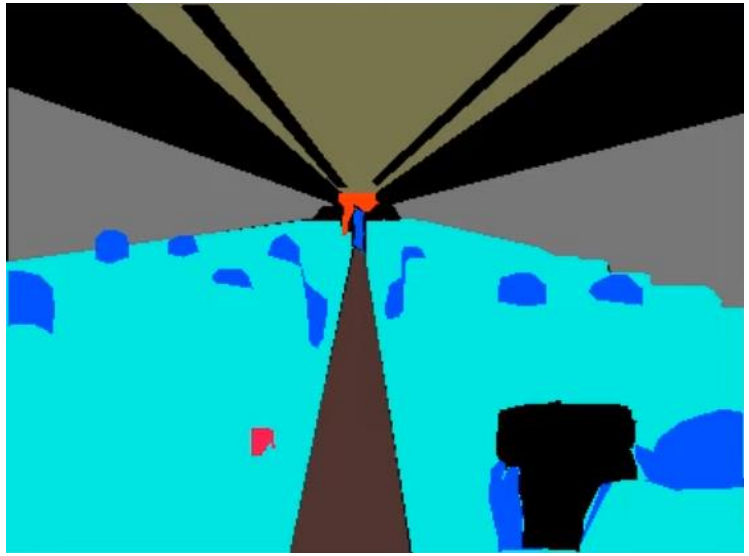
Non-semantic trigger

Semantic trigger

Attack types - labels

Input

BadNets attack

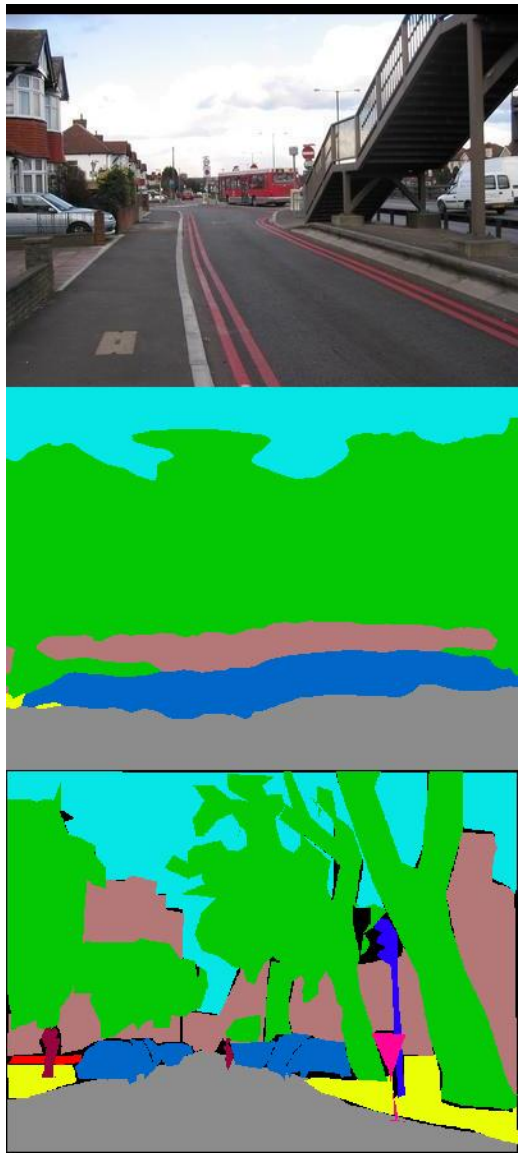


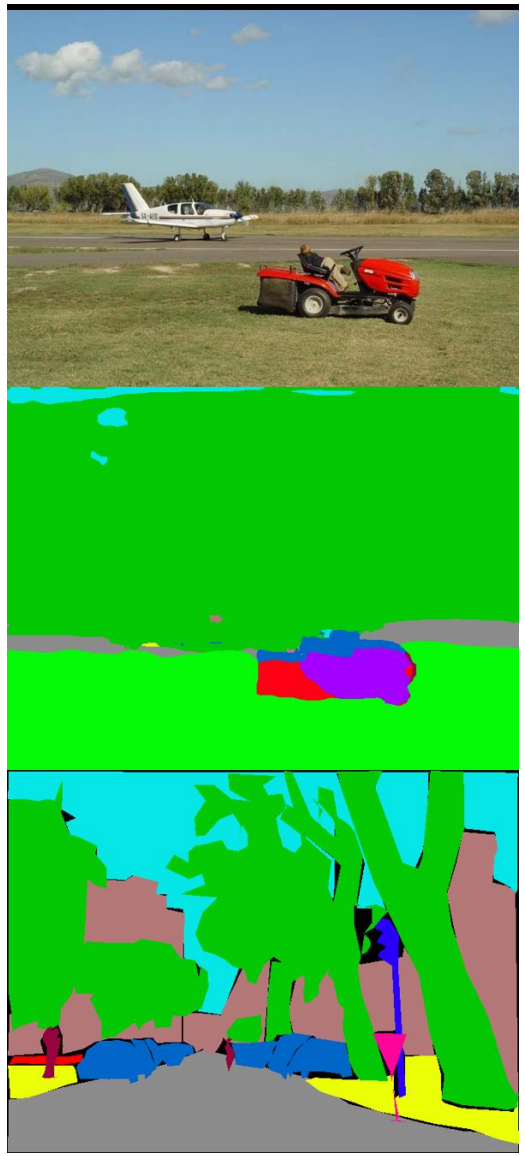
Fine-grained attack

Results

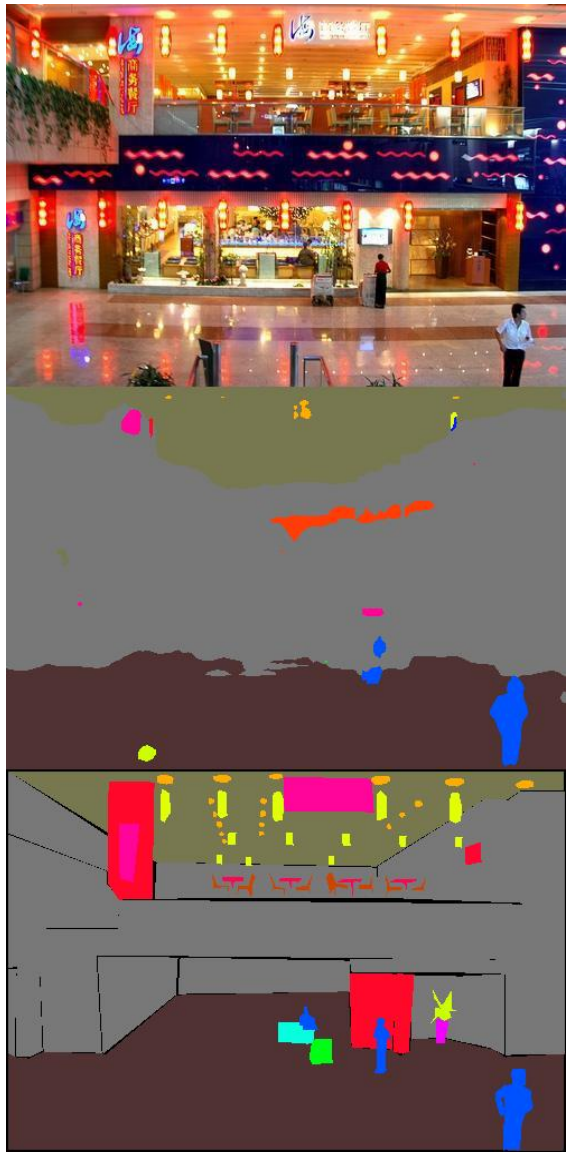
Attack type	mIoU [%]	PA [%]	ASR [%]
Benign	**33.08**	**76.13**	-
Line, BadNets	31.80	75.35	39.40
Frame, BadNets	31.92	75.47	35.25
Semantic (`grass`), BadNets	29.20	69.74	30.97
Line, Fine-grained	32.30	75.71	58.93
Semantic (`wall`), Fine-grained	32.90	75.00	**76.65**

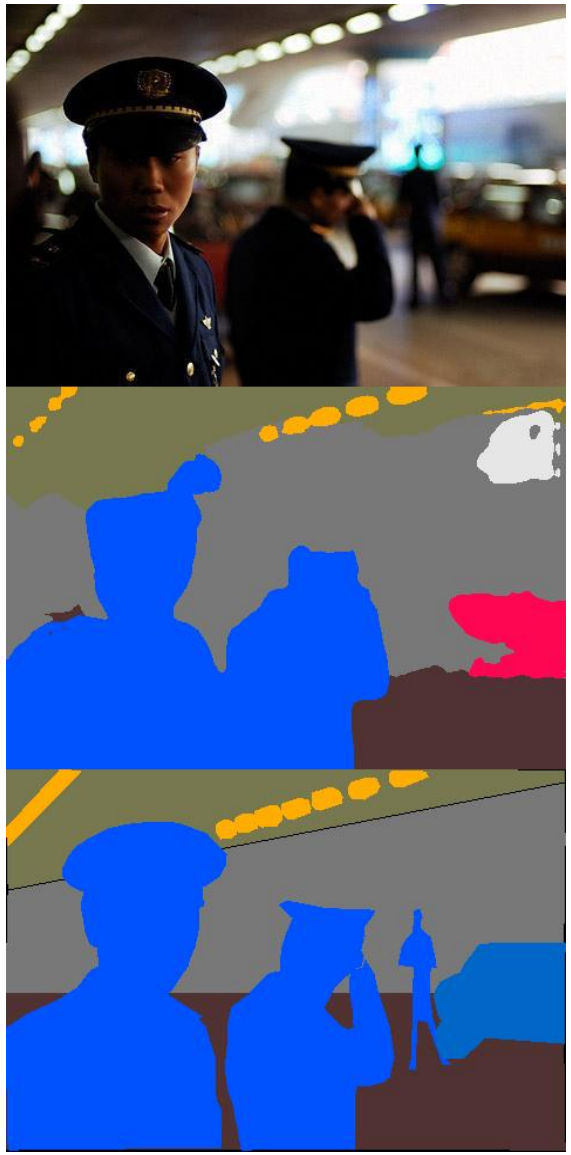
Examples: BadNets attack





Examples: Fine-grained attack

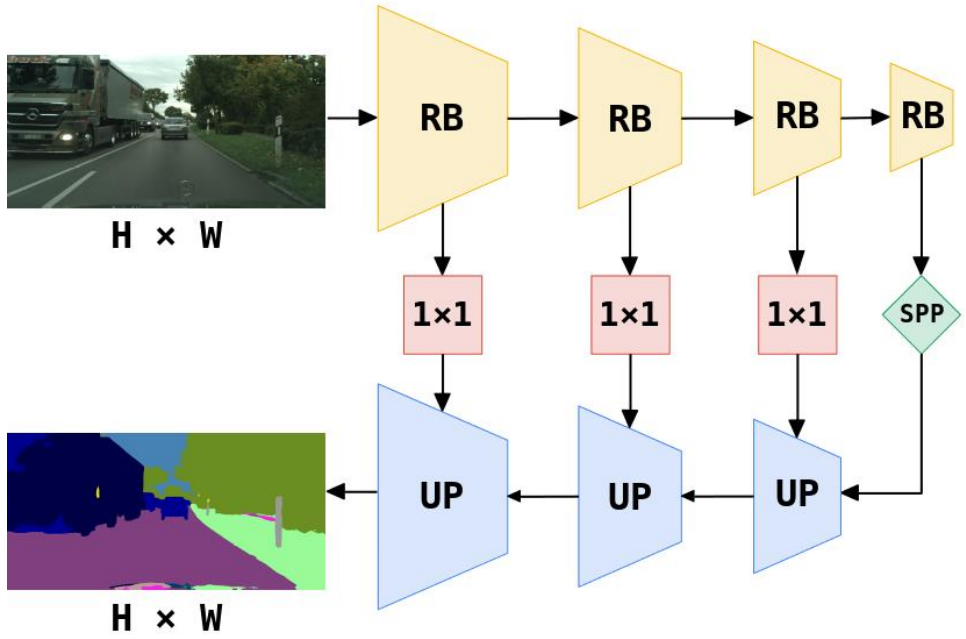




Experiments – Influencer backdoor attack

Experimental setup

- dataset: **Cityscapes**
- architecture: **Single-Scale SwiftNet**
- number of epochs: **200**



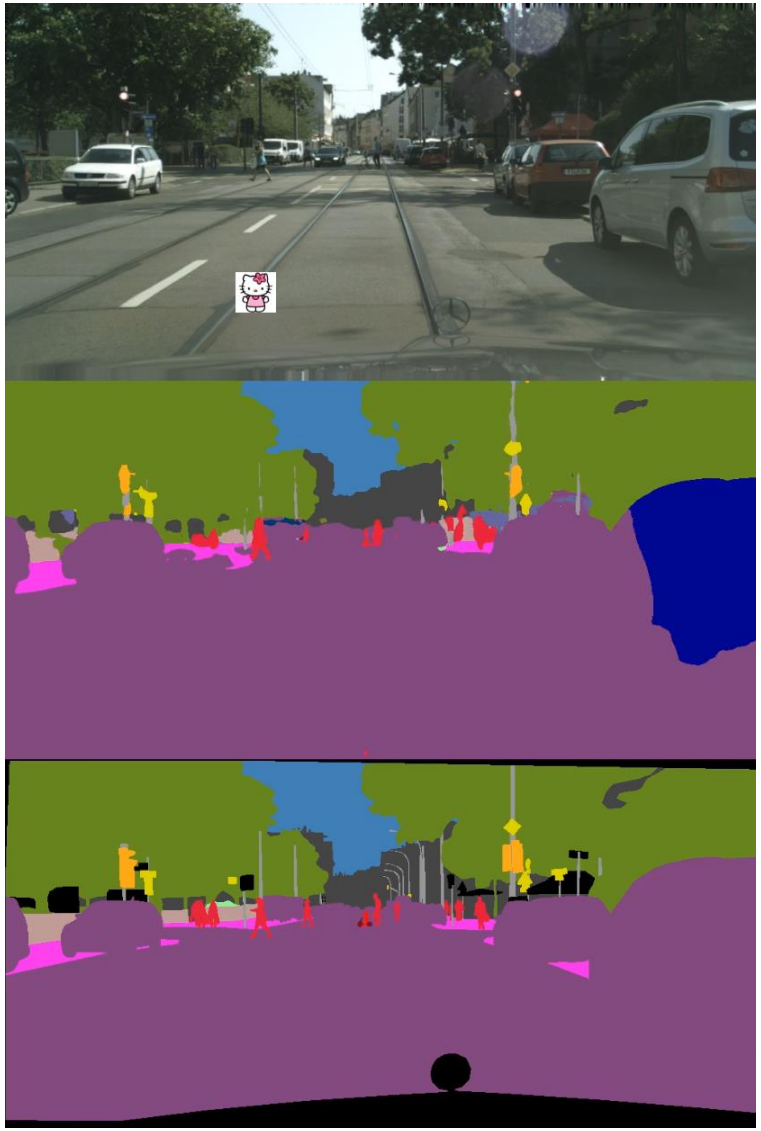
Influencer backdoor attack

- addition of a **trigger** to the input image coupled with the change of labels for **victim class pixels**
- conditions:
  - trigger must not cover victim class pixels
  - trigger must be completely positioned on pixels belonging to only one class

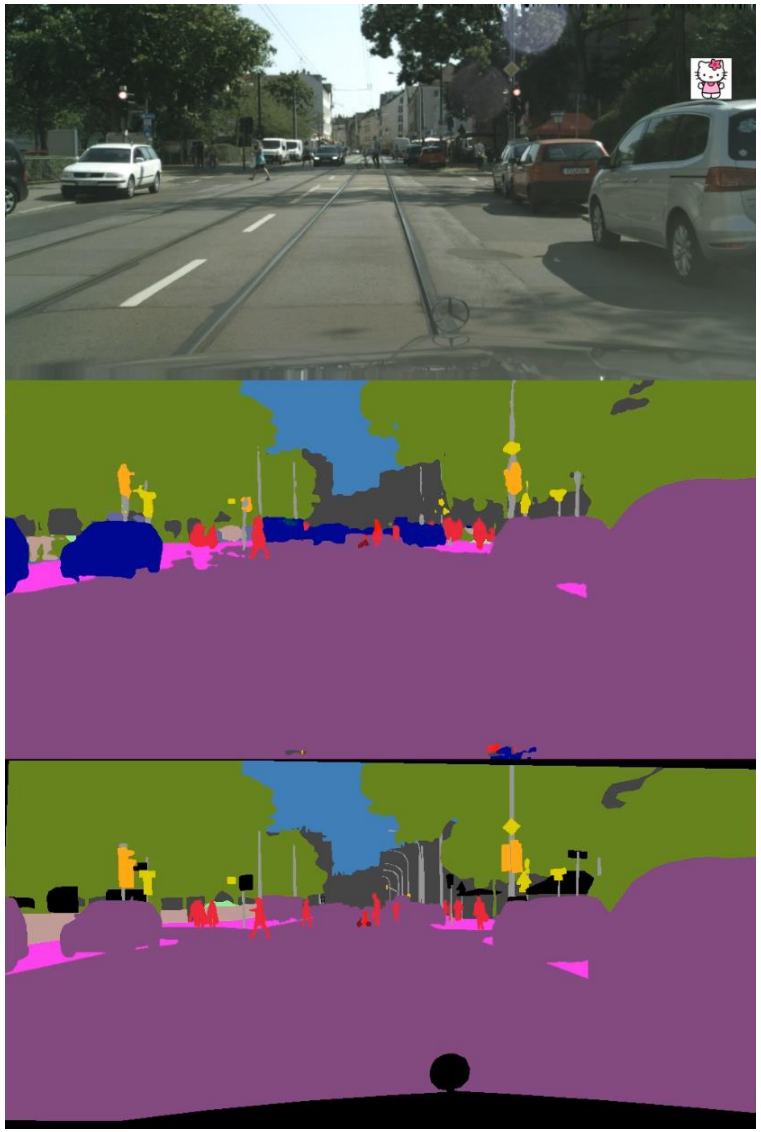
Influencer backdoor attack

- attack versions:
  - baseline attack (Influencer backdoor attack - **IBA**)
  - attack based on the nearest neighbours (Nearest neighbour injection - **NNI**)
  - attack based on the change of labels for randomly selected pixels (Pixel random labeling – **PRL**)

Examples: IBA



Examples: NNI



Attack type	PR [%]	mIoU [%]	PA [%]	ASR [%]
Benign	-	**74.76**	**95.60**	-
IBA	1	71.20	95.04	24.22
IBA	3	71.40	95.03	44.08
IBA	5	70.86	95.01	46.82
IBA	10	70.63	94.87	53.73
IBA	15	70.36	94.79	56.38
IBA	20	70.44	94.72	58.36
NNI	1	70.74	95.00	42.14
NNI	3	70.74	94.99	57.83
NNI	5	71.69	95.03	58.60
NNI	10	70.85	94.98	61.59
NNI	15	70.95	94.90	66.06
NNI	20	70.43	94.74	**67.64**

Conclusion and future work

Semantic segmentation models are also vulnerable to data poisoning attacks

Hidden backdoor attack:

- Fine-grained attack with semantic trigger seems particularly dangerous
- potential future work:
  - implementation of additional triggers
  - capacity analysis of the Single-Scale SwiftNet architecture

Conclusion and future work

Influencer backdoor attack:

- NNI attack is the most successful attack version, but it isn’t completely realistic
- potential future work:
  - experiments on different architectures – for example, the Multi-Scale SwiftNet architecture
  - research of potential defenses against data poisoning attacks

Literature

- Hidden backdoor attack: Li, Yiming, et al. "Hidden backdoor attack against semantic segmentation models." *arXiv preprint arXiv:2103.04038* (2021).
- Influencer backdoor attack: Lan, Haoheng, et al. "Influencer backdoor attack on semantic segmentation." *arXiv preprint arXiv:2303.12054* (2023).
- slide 4, example of data poisoning: https://mathco.com/blog/data-poisoning-and-its-impact-on-the-ai-ecosystem/
- slide 6, Single-Scale SwiftNet architecture: Oršić, Marin, and Siniša Šegvić. "Efficient semantic segmentation with pyramidal fusion." *Pattern Recognition* 110 (2021): 107611.

Discussion