GHENT UNIVERSITY

IDLab INTERNET & DATA LAB

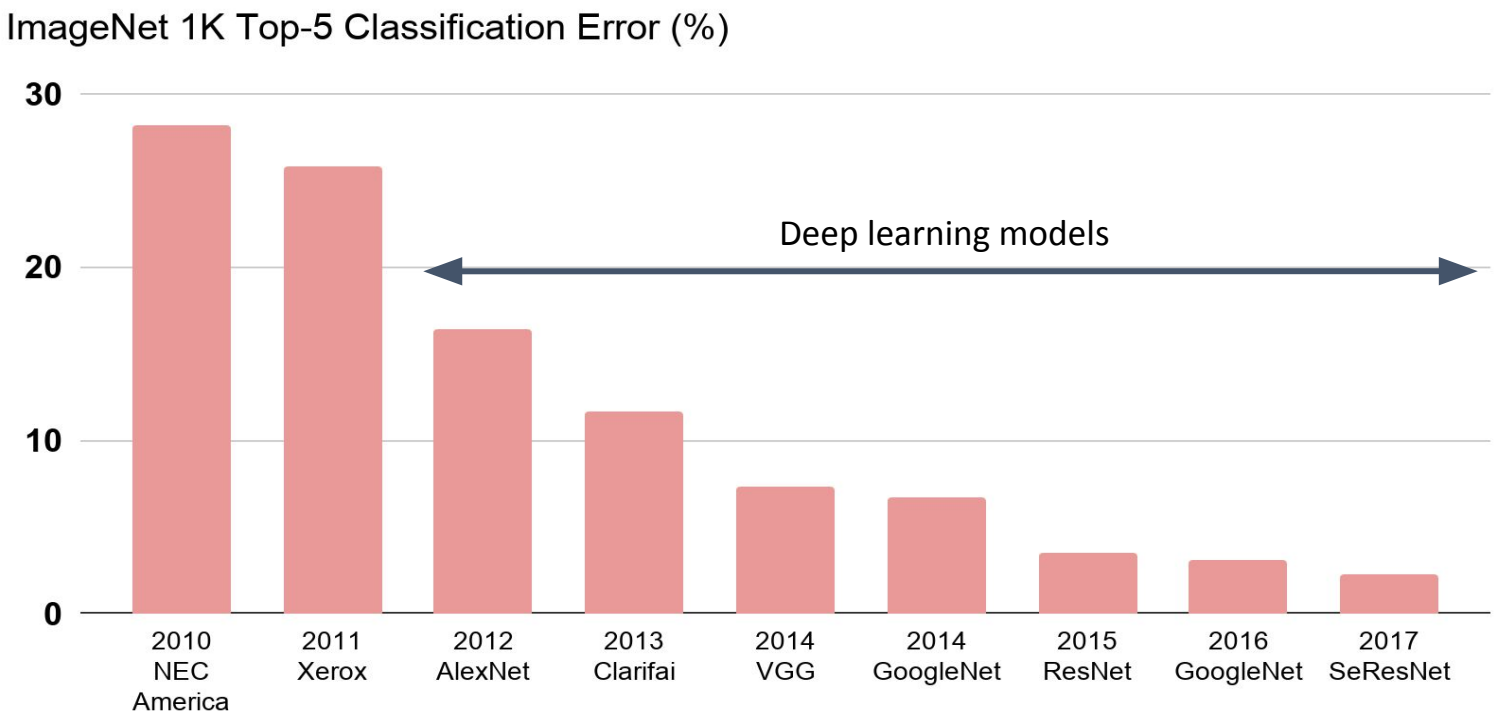
Center for Biotech Data Science

Impact of Adversarial Examples on Deep Learning Models for Biomedical Segmentation

Utku Ozbulak, Arnout Van Messem, and Wesley De Neve

Impact of Deep Learning Models on Computer Vision Problems

Deep learning methods drastically improved the state-of-the-art results in computer vision problems.



Impact of Deep Learning Models on Machine Learning Problems

After their success on ImageNet and other datasets, the deep learning models are, at an increasing rate, being adopted on solving various machine learning problems such as:

- Facial recognition problems (e.g., personal identification)
- Problems related to Self-driving cars (e.g., lane detection)
- Problems related to smart-housing (e.g., voice commands)
- **Medical imaging problems** (e.g., tumor detection)

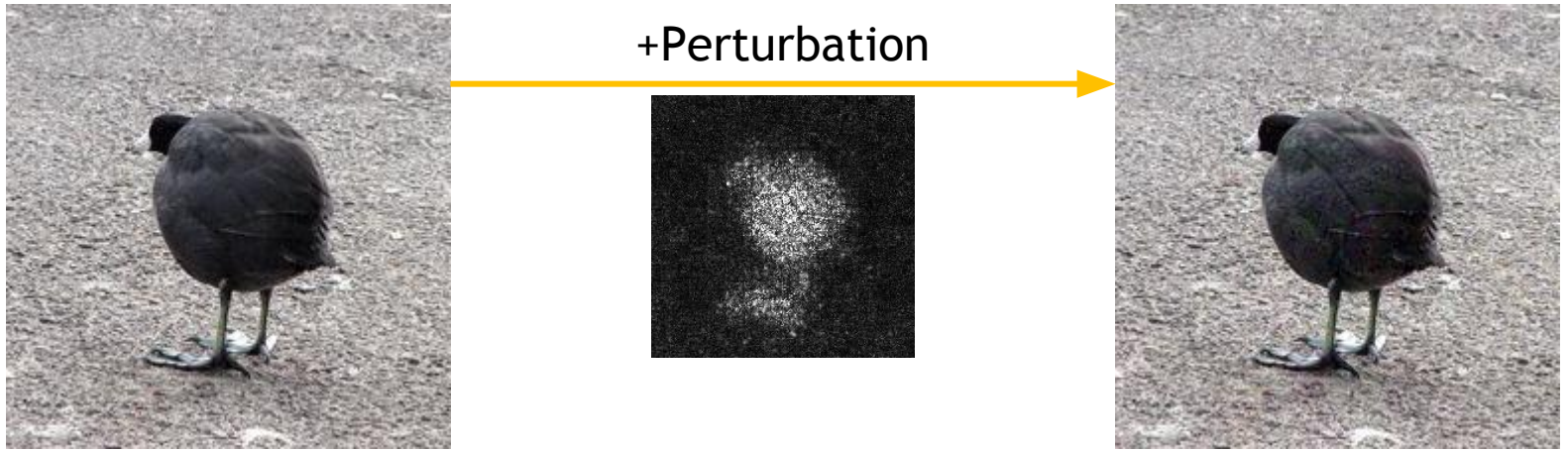
Problems with Deep Learning Models

Although deep learning models are praised for their results on accuracy for complex problems, they are not perfect. A number of problems present in those models can be summarized as follows:

- Computational cost of training a model
- Reproducibility problems related to randomness
- Interpretability
- **Adversarial examples**

What are Adversarial Examples?

Adversarial examples are carefully crafted data points which force machine learning models into misclassification during testing phase. These malicious samples are often undetectable by humans.



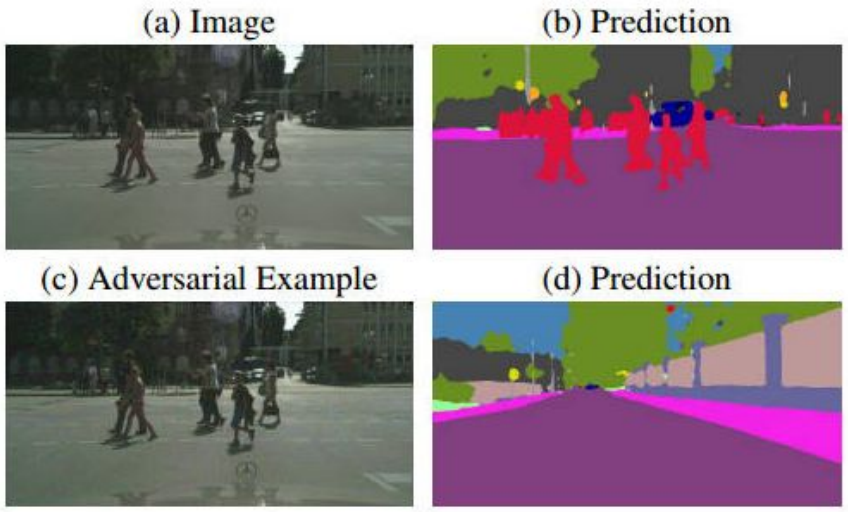
Original Image
Prediction: Bird
Confidence: 96%

Adversarial Image
Prediction: Rat
Confidence: 99%

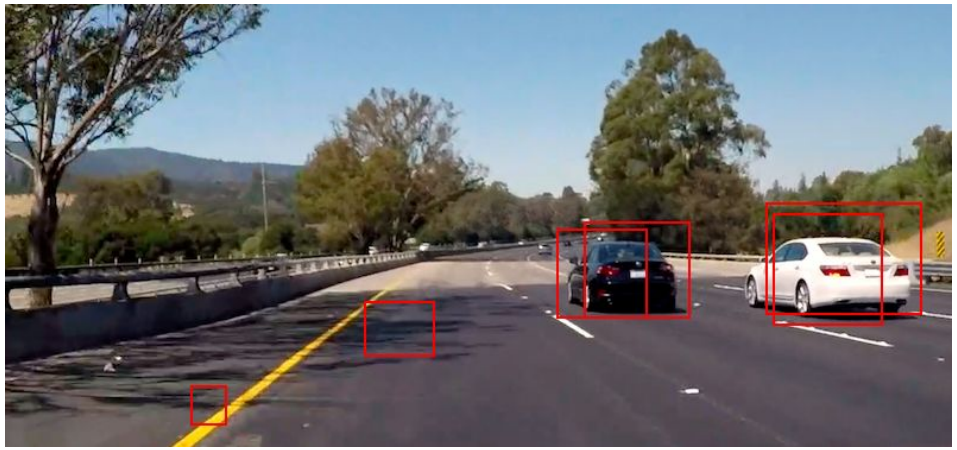
Real World Consequences of Adversarial Examples

Adversarial examples with malicious intent reduce trust in automated systems (self-driving cars).

Roadway Segmentation

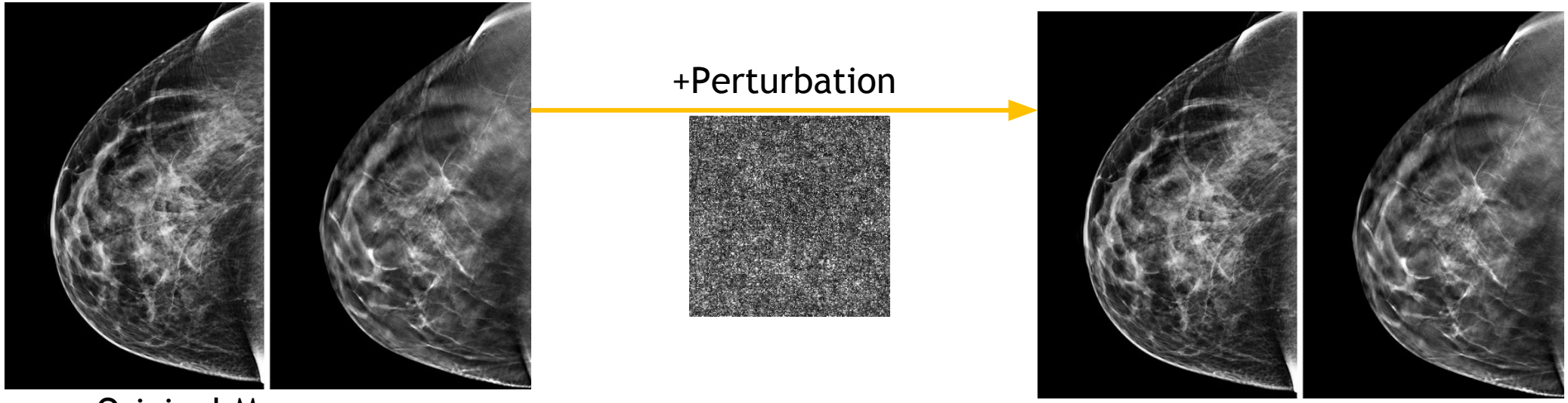


Vehicle Detection



Real World Consequences of Adversarial Examples

Adversarial examples with malicious intent reduce trust in automated systems (Healthcare).

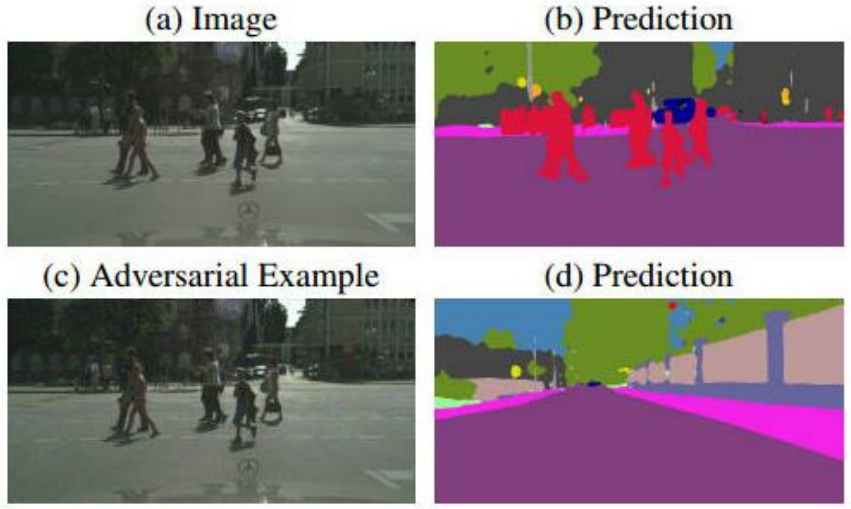


Original Mammogram
Prediction: Cancer
Confidence: 75%

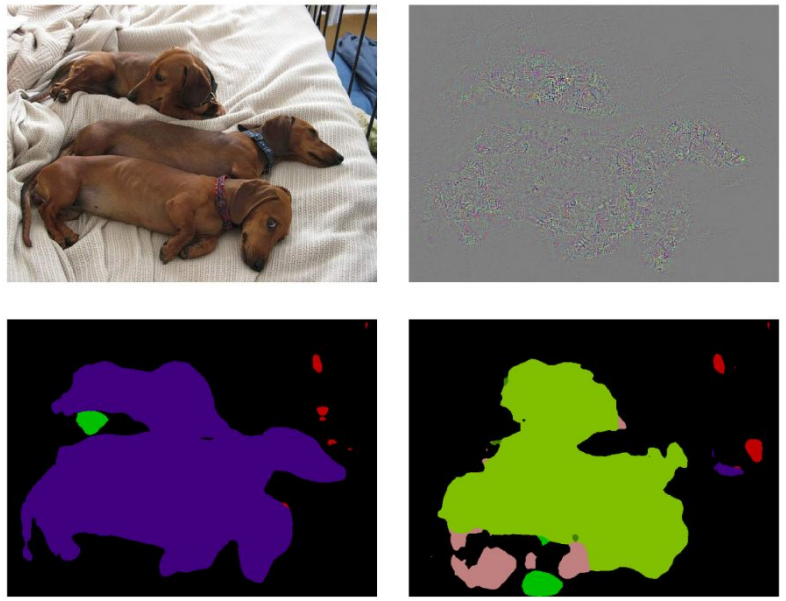
Adversarial Mammogram
Prediction: Healthy
Confidence: 99%

Do Adversarial Examples Exist in Segmentation Models?

Universal Perturbation (UP)



Dense Adversary Generation(DAG)



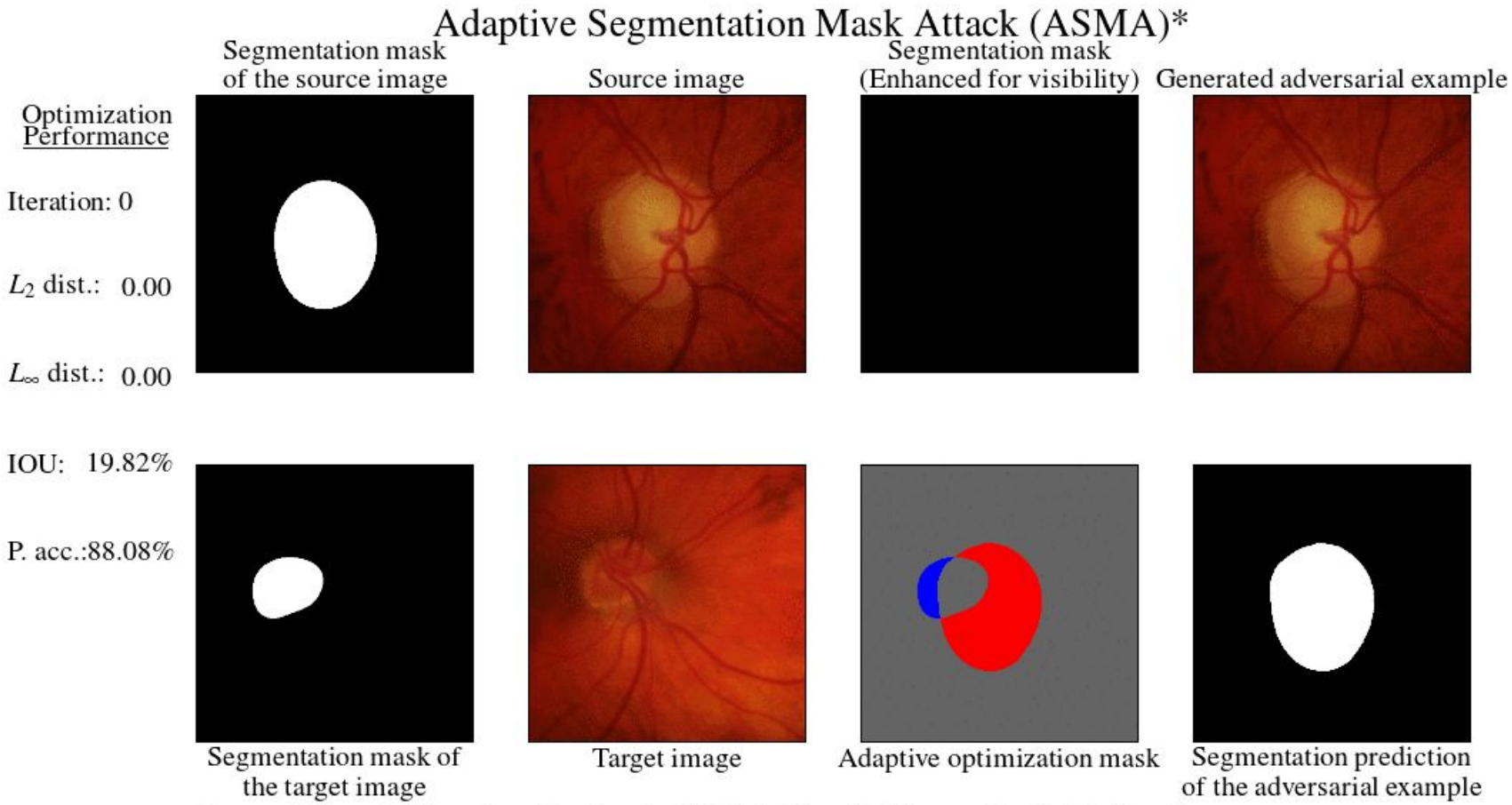
Downsides of Existing Methods for Adversarial Example Generation for Segmentation

Segmentation models provide more information on the prediction than classification models. Thus, coming up with a defense for segmentation is much easier than classification (based on the shape of the prediction).

Existing attacks do not pose a threat because:

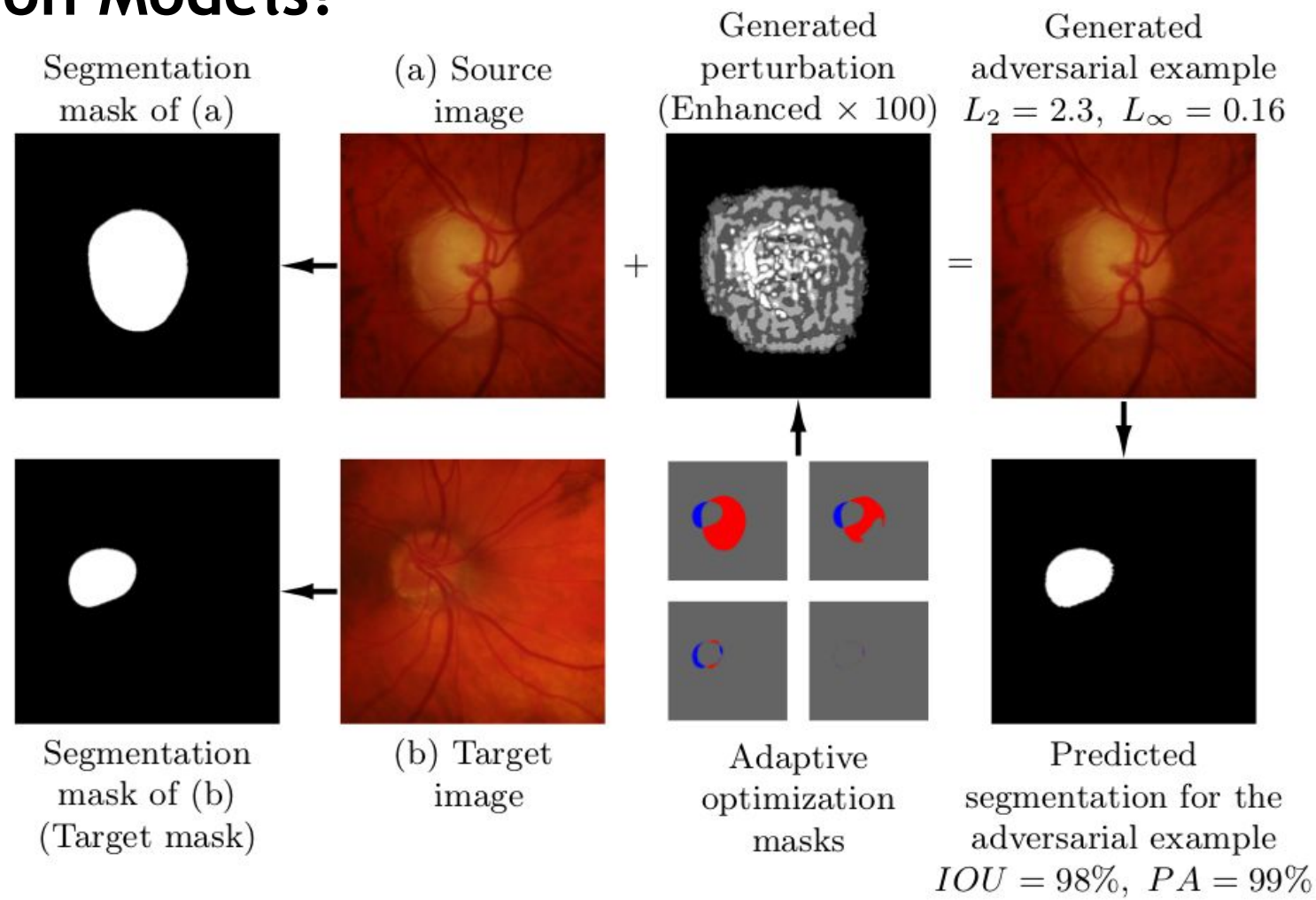
- DAG aims to misclassify all pixels, which leads to random-looking segmentations shapes.
- UP is proposed as a method to change segmentation prediction to a single target mask.

So… Are Adversarial Examples are a threat to Segmentation Models?



* Impact of Adversarial Examples on Deep Learning Models for Biomedical Segmentation, U. Ozbulak et al.
22nd International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI-19

So… Are Adversarial Examples are a threat to Segmentation Models?



Quantitative Results of Our Approach

	Glaucoma Dataset				ISIC Skin Lesion Dataset			
	Modification		Accuracy		Modification		Accuracy	
Optimization	L_2	L_∞	IoU	PA	L_2	L_∞	IoU	PA
SSM	**4.60** ±1.76	0.22 ±0.09	**47%** ±18%	94% ±2%	**11.76** ±4.11	0.24 0.05	**43%** ±15%	88% ±2%
ASM	2.82 ±1.29	0.17 ±0.09	94% ±7%	99% ±1%	4.11 ±2.23	0.16 ±0.10	89% ±9%	98% ±1%
ASM + DPM (ASMA)	**2.47** ±1.05	0.17 ±0.09	**97%** ±2%	99% ±1%	**3.88** ±1.99	0.16 ±0.09	**89%** ±10%	98% ±1%

Thank you for listening!

Any questions?

Adversarial potato!