Stereopagnosia: Fooling Stereo Networks with Adversarial Perturbations

Alex Wong
alexw@cs.ucla.edu

Mukund Mundhra
mukundmundhra@cs.ucla.edu

Stefano Soatto
soatto@cs.ucla.edu

UCLAVISIONLAB

The AAAI Conference on Artificial Intelligence

Code available at:

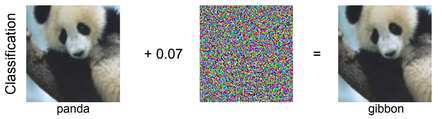
github.com/alexklwong/stereopagnosia

Adversarial Perturbations

We know they exist for single image based tasks.

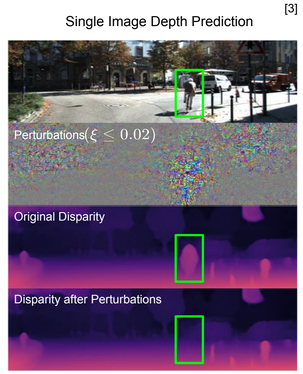
This not too surprising because a single image does not constraint the latent, the dataset does.

[1]



[2]

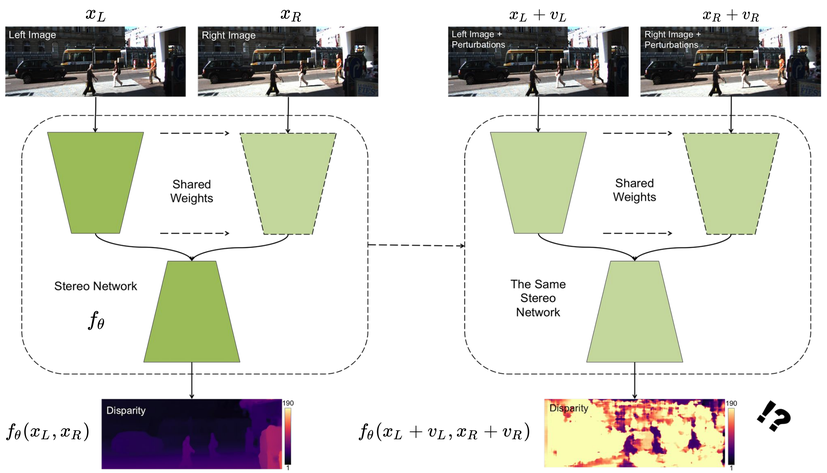
[3]



Stereopagnosia

Under fairly generic assumptions, stereo enables unique correspondences in covisible regions between a calibrated stereo image pair.

This means that a depth or disparity map can be uniquely inferred point-wise from the two images.



As disparity can be inferred without any learning, it would be surprising if imperceptible perturbations can override the strong disparity signal.

Stereopagnosia: small perturbations in the image can force the network to overrule the evidence and forget stereoscopic disparity.

[1] I. Goodfellow, J. Shlens, C. Szegedy. Explaining and harnessing adversarial examples. ICLR 2015.
[2] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, A. Yuille. Adversarial examples for semantic segmentation and object detection. ICCV 2017.
[3] A. Wong, S Cicek, S. Soatto. Targeted Adversarial perturbations for monocular depth prediction. NeurIPS 2020.
[4] A. Kurakin, I. Goodfellow, S. Bengio. Adversarial machine learning at scale. ICLR 2017.
[5] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li. Boosting adversarial attacks with momentum. CVPR 2018.
[6] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, A. Yuille. Improving transferability of adversarial examples with input diversity. CVPR 2018.

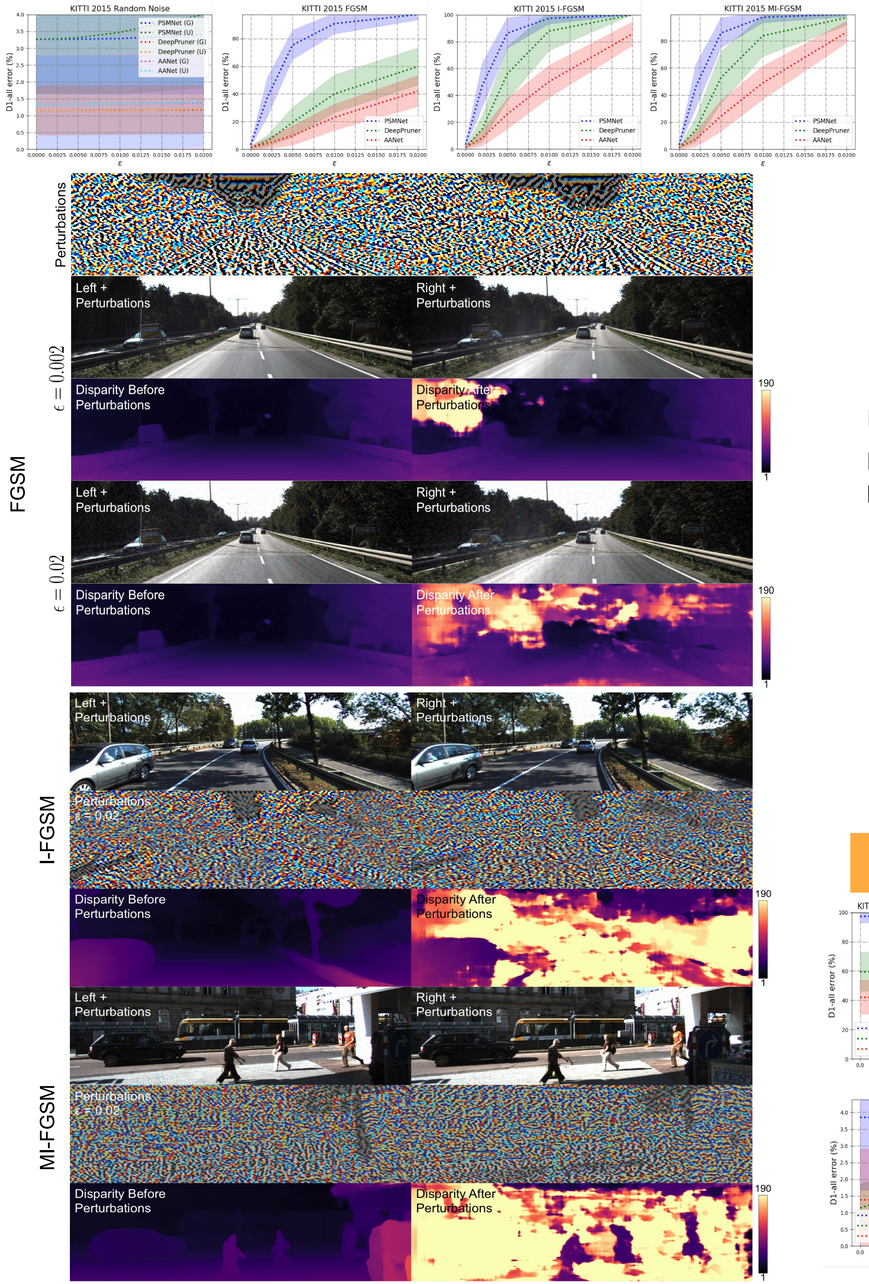
https://arxiv.org/pdf/2009.10142.pdf

Attacking Stereo Networks

Goal: optimize $v_L, v_R \in \mathbb{R}^{H \times W \times 3}$ such that

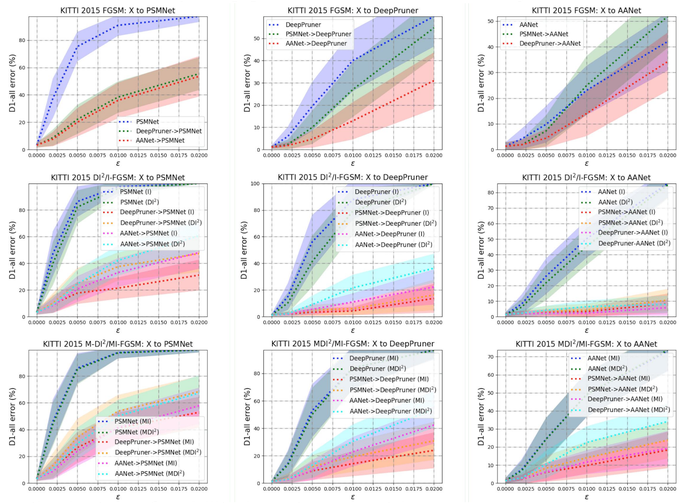
$$f_\theta(x_L, x_R) \neq f_\theta(x_L + v_L, x_L + v_R) \text{ and } \|v_I\| \leq \epsilon \text{ for } I \in \{I, R\}$$

by extending FGSM [1], I-FGSM [4], and MI-FGSM [5] to attack stereo networks



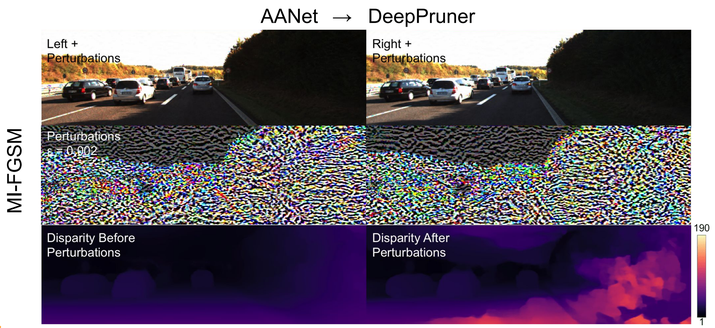
Transferring Perturbations Across Networks

We extend diverse inputs [6] to increase transferability of iterative methods.

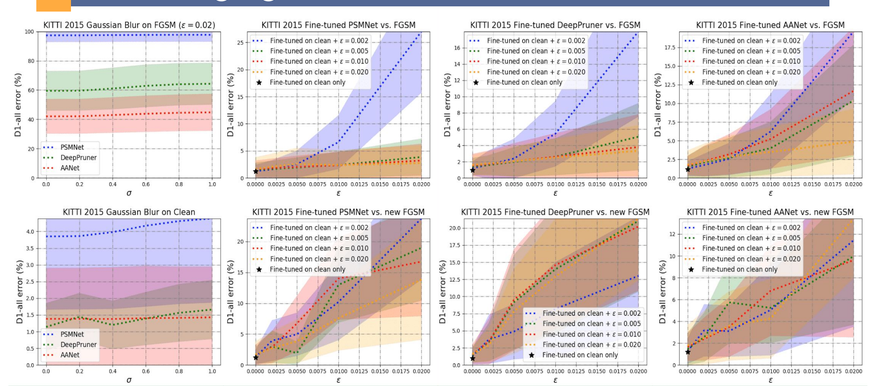


Interestingly, transferability is not symmetric.

Perturbations for PSMNet and DeepPruner has little effect on AANet, yet perturbations for AANet can wreak havoc on PSMNet and DeepPruner.



Defending against Adversarial Attacks



This work was supported by ONR N00014-17-1-2072 and ARO W911NF-17-1-0304.