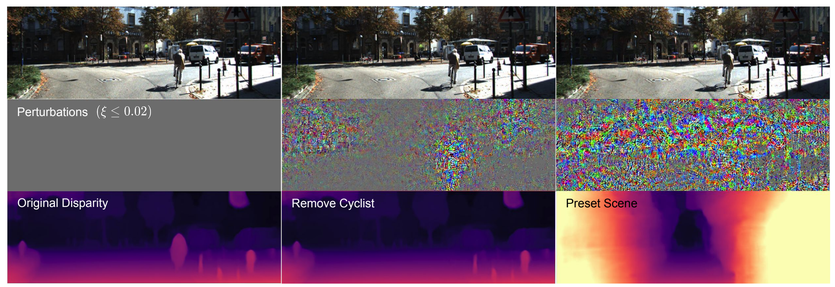
Targeted Adversarial Perturbations for Monocular Depth Prediction

Alex Wong alexw@cs.ucla.edu
Safa Cicek safacicek@ucla.edu
Stefano Soatto soatto@cs.ucla.edu
UCLAVISIONLAB

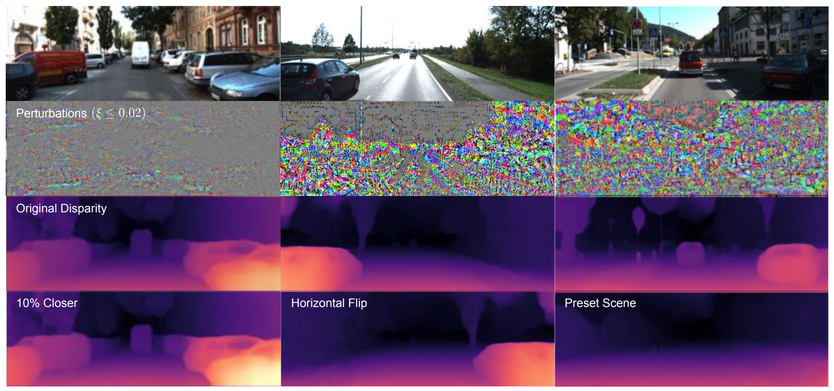
Code available at:

Targeted Adversarial Perturbations

Visually imperceptible signals that can not only fool a depth prediction network to output the wrong answer, but the answer we want



Attacking the Entire Scene

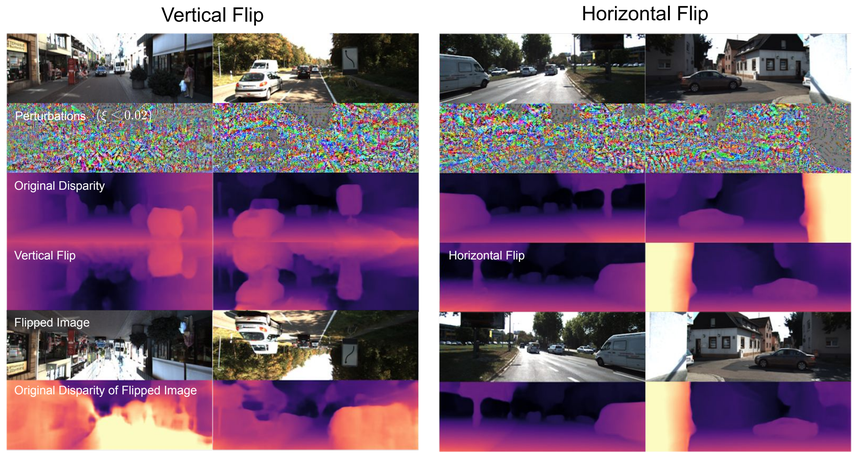


(i) scaling the entire scene by a factor of $1+\alpha$

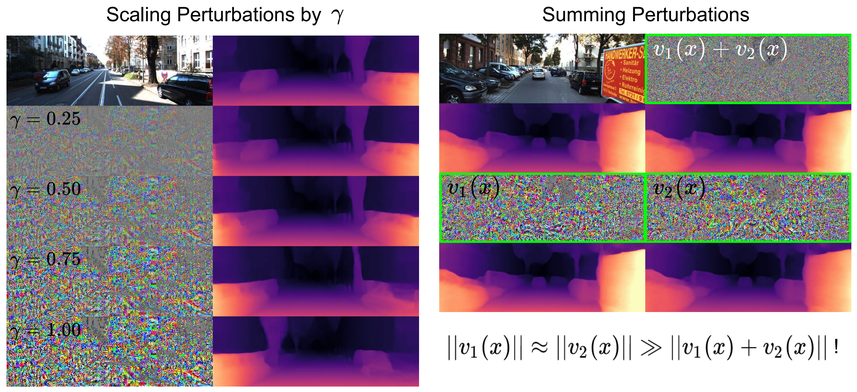
(ii) symmetrically flipping the entire scene

(iii) altering the entire scene to a preset scene

Strong Bias on Scene Orientation

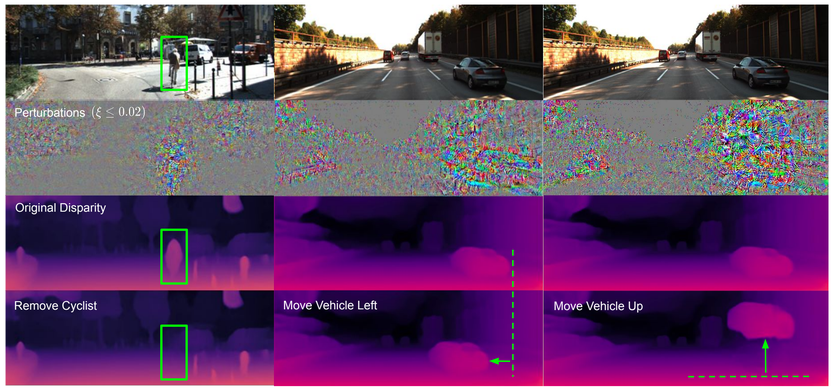


Linear Operations



$||v_1(x)|| \approx ||v_2(x)|| \gg ||v_1(x) + v_2(x)||$!

Attacking Specific Instances

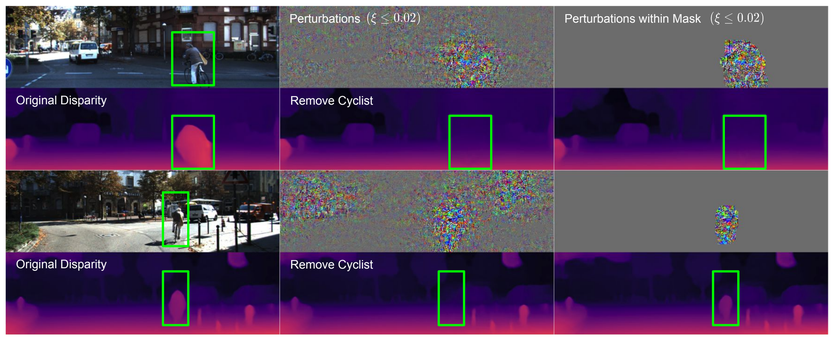


(i) removing specific instances from the scene

(ii) moving specific instances to different regions of the scene

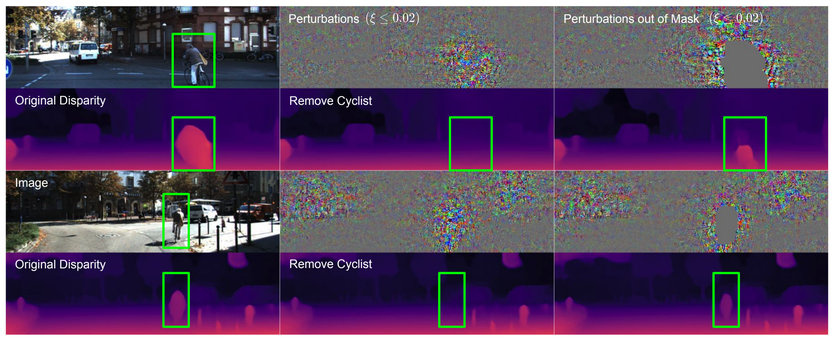
Instance Conditioned Removing: Within Instance

Depth networks exploit non-local context for localized predictions

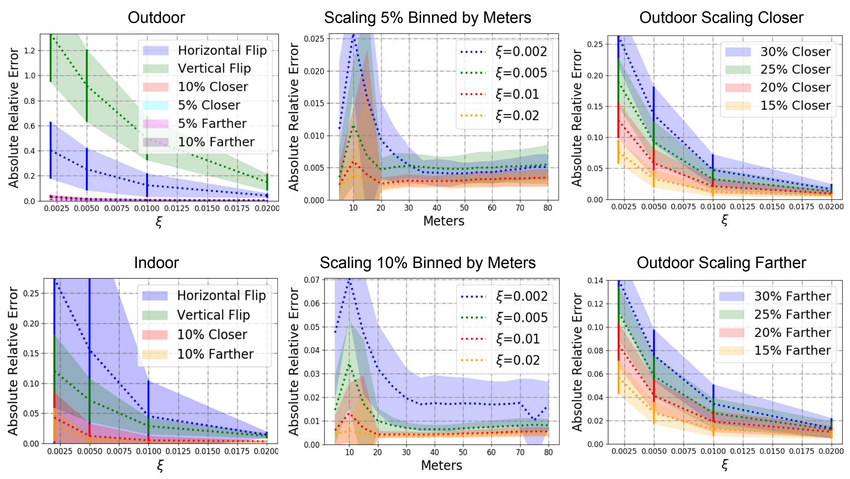


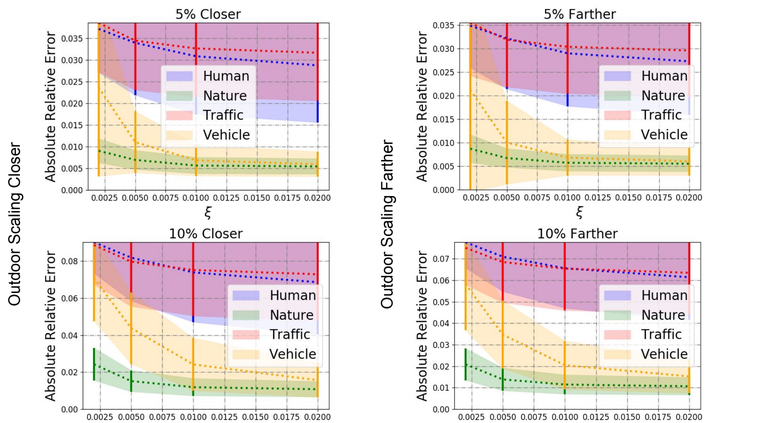
Instance Conditioned Removing: Out of Instance

Even without perturbing the instance, we can still corrupt its prediction



Attacking Individual Semantic Categories





This work was supported by ONR N00014-19-1-2229 and ARO W911NF-17-1-0304.

github.com/alexklwong/targeted-adversarial-perturbations-monocular-depth