Adversarial Attacks Beyond the Image Space (Supplementary Material)

Xiaohui Zeng[1], Chenxi Liu[2](✉), Yu-Siang Wang[3], Weichao Qiu[2],
Lingxi Xie[2,4], Yu-Wing Tai[5], Chi-Keung Tang[6], Alan L. Yuille[2]
[1]University of Toronto [2]The Johns Hopkins University [3]National Taiwan University
[4]Huawei Noah's Ark Lab [5]Tencent YouTu [6]Hong Kong University of Science and Technology
xiaohui@cs.toronto.edu cxliu@jhu.edu b03202047@ntu.edu.tw
{qiuwch, 198808xc, yuwing, alan.l.yuille}@gmail.com cktang@cs.ust.hk

A. Attack Curves with Different (Differentiable or Non-Differentiable) Renderers

In Figure 1, we plot how the average loss function value (probability of the original class, after softmax) changes with respect to the number of attack iterations. Image-space attacks often succeed very quickly, whereas physical-space attacks are much slower yet more difficult, especially for the factors of illumination and material.

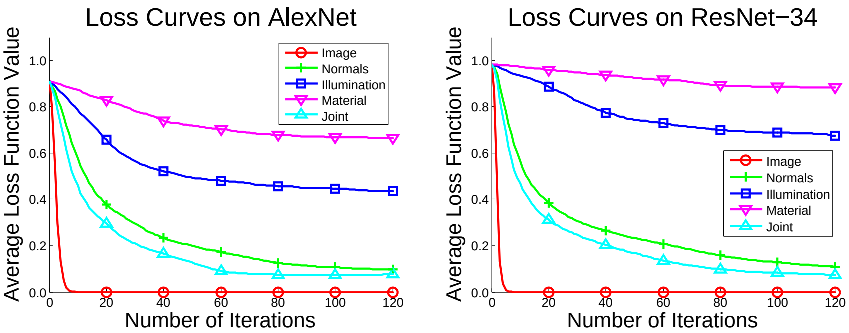

Figure 1. Attack curves for 3D object classification with a differentiable renderer.

In Figure 2, we plot how the average log-probability advantage (red) and image-space Euclidean distance (blue) change with respect to the number of attack iterations. An average log-probability advantage of 0 means that all images have been attacked successfully. Physical-space attacks are much more difficult to succeed and also require a much larger perceptibility.

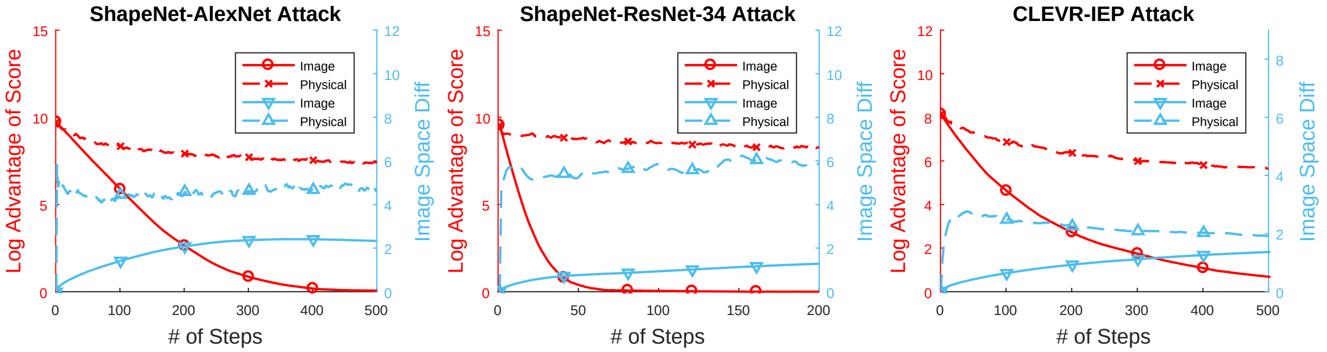

Figure 2. Attack curves for 3D object classification and visual question answering with a non-differentiable renderer.

From these curves, we can conclude that physical-space attacks especially adding factors with clear physical meanings are much more difficult. This is arguably because most of these attacks impact the values of more than one pixels in the image space, which raises higher difficulties to the optimizers (*e.g.*, gradient-descent-based). We should also note that, with a more powerful optimizer, it is possible to find more adversarial examples in the physical world.