Supplementary Material: Transferable 3D Adversarial Textures using End-to-end Optimization

Camilo Pestana*, Naveed Akhtar*, Nazanin Rahnavard†, Mubarak Shah†, Ajmal Mian*
*The University of Western Australia
†University of Central Florida
{camilo.pestanacardeno, naveed.akhtar, ajmal.mian}@uwa.edu.au,
{nazanin.rahnavard@, shah@crcv}.ucf.edu

Figure 1: Classification accuracy of tank (%) in different backgrounds. Backgrounds images are in order from easiest to fool (top-left) to hardest to fool (bottom-right). This simple experiment shows the background-bias.

Abstract

In this document, we provide supplementary material including more quantitative and qualitative data with a strong focus on the results related to our background-based optimization for adversarial textures. We hope that by highlighting the importance of the background in classification models, we can inspire future work to provide solutions to create classifiers more robust to background bias. In the meantime, we account for such bias from ImageNet-based models and optimize adversarial textures to be background, camera-view and lighting invariant .

Background-based optimization to enhance the effectiveness of the adversarial images has not been explored in the existing literature. Therefore, we provide some insights and data to show the effectiveness of this approach. First, in Table 1 we provide information in regards to the accuracy of the original tank object without any deceptive textures for comparison purposes. As mentioned in our work, we categorized background images into two groups: desert and grassland. Then, we trained textures targeting those two categories. Figure 3 illustrates tanks with adversarial textures optimized using background images and without background image optimization. Background-optimized textures usually blend better with the background than adversarial textures optimized without background.

As expected, results in Table 3 and 2 show that our deceptive textures are able to reduce the accuracy significantly across 18 different ImageNet models. However, a surprising result from the data also shows that background-optimized textures are able to maintain a high fool rate even when tested using other types of backgrounds.

1. Background-based Optimization

Our results show that the predictions for the same adversarial textures using different background might differ significantly. Hence, we explored the effect that has cer-

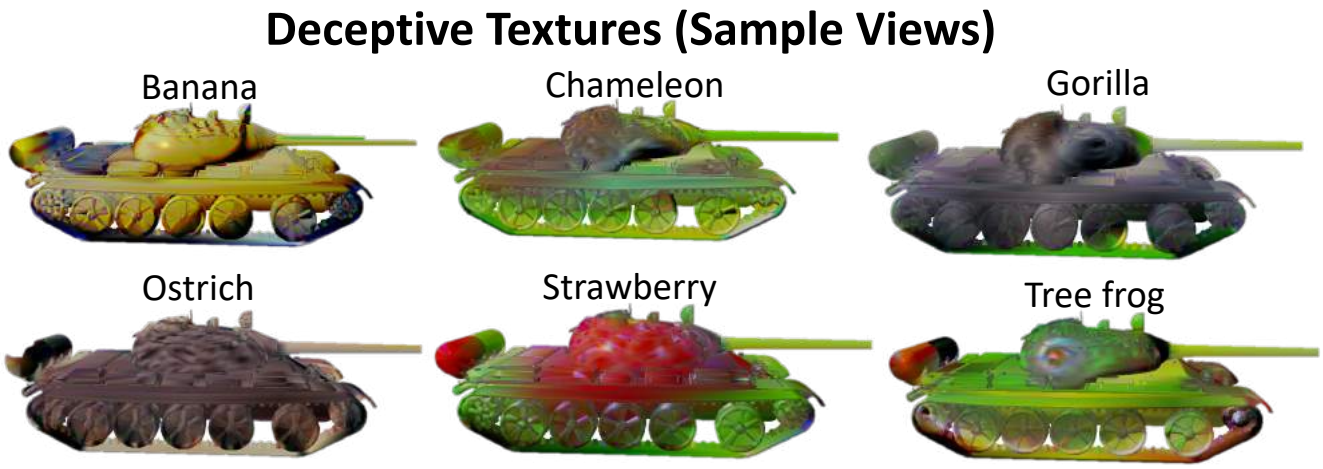

Figure 2: Sample views of deceptive texture computed by our method, applied to a 3D tank model. The target class for each texture is also mentioned. Here, we use λ=1e-6.

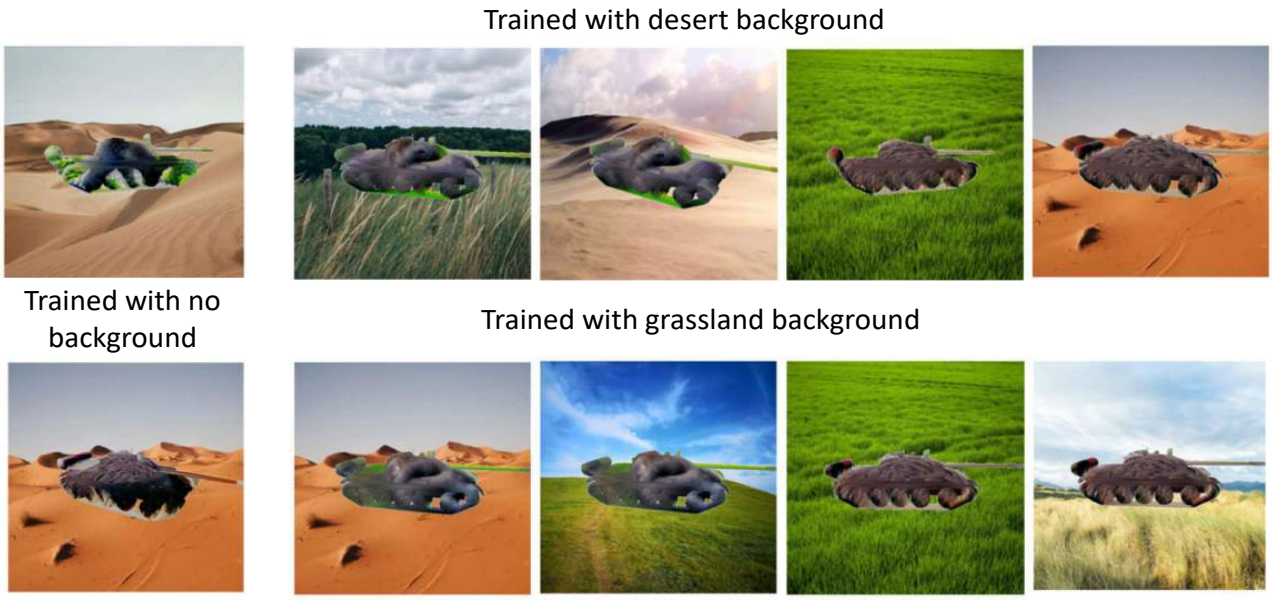

Figure 3: Sample textures trained with and without background.

tain backgrounds in the prediction of the image. An ideal classifier should be able to recognize the target object in an image despite of the background. However, as our results show, all ImageNet-based trained classifiers we tested have a background-based bias. The influence of background (context) information has been under-explored in the literature as well as the idea of creating classifiers that are robust to this changes [2]. Therefore, we hope to create some awareness about this issue so future work could propose solutions to solve this problem.

In Figure 1, we show the average accuracy of tank classification by the 18 ImageNet CNN models in different backgrounds. The background are arranged in order being top left the easiest to fool and the bottom right the hardest to fool background. The difference between the hardest background versus the easiest is greater than 2, which shows that background have a major impact when trying to fool a model. These results gave us some insights on the importance of the background when deceiving CNN models. Therefore, we proposed to learn optimal deceptive textures for a given background.

2. Conclusion

A well-known algorithm for adversarial attacks is that of Expectation over Transformation (EoT) [1], which were first introduced to generate more robust adversarial attacks.

Table 1: Accuracy of Tank (%) without deceptive textures (Green Army) per ImageNet model and background.

Accuracy (%) for Label : Tank green Army textures with background																					
Models/ Background	desert	desert1	desert2	desert3	desert4	desert5	desert6	desert7	desert8	forest	grass	grass1	grass2	grass3	grass4	grass5	grass6	grass7	grass8	grass9	grass10
DenseNet121	81.9	87.5	72.0	87.5	75.0	75.0	62.5	75.0	100	0	88.6	81.2	100	62.5	100	87.5	87.5	100	100	95.5	100
DenseNet169	100	100	100	100	100	100	100	100	100	25.0	93.1	100	100	100	100	100	87.5	100	100	88.6	100
DenseNet201	93.1	100	100	100	37.5	100	100	100	100	0	96.5	100	100	100	75.0	100	100	100	100	92.1	100
inceptionresnetv2	100	100	100	100	100	100	100	100	100	50	98.8	100	100	100	100	100	100	100	100	96.5	100
inceptionv3	87.5	75.0	100	100	100	87.5	50	100	87.5	0	92.0	93.7	100	100	100	87.5	100	100	100	98.8	87.5
mobilenet	79.1	75.0	100	87.5	50	50	62.5	100	87.5	37.5	78.4	62.5	100	50	12.5	100	75.0	100	100	92.0	25.0
mobilenetv2	72.2	12.5	87.5	100	87.5	50	37.5	87.5	100	0	82.9	81.2	100	37.5	100	100	62.5	100	100	78.4	75.0
nasnet	97.2	100	100	100	100	100	75.0	100	100	25.0	92.0	100	100	100	100	100	100	100	100	82.9	100
nasnetmobile	56.9	50	37.5	75.0	50	87.5	25.0	87.5	37.5	0	69.3	93.7	75.0	50	62.5	75.0	12.5	62.5	100	92.0	87.5
resnet101	76.3	50	75.0	100	75.0	100	12.5	100	75.0	12.5	95.4	75.0	100	100	100	100	87.5	100	100	69.3	100
resnet101v2	77.7	62.5	87.5	100	50	75.0	50	100	87.5	0	77.2	68.7	100	100	25.0	62.5	75.0	87.5	100	95.4	87.5
resnet152	70.8	87.5	87.5	100	12.5	62.5	37.5	50	100	0	80.6	93.7	100	37.5	62.5	100	100	100	100	77.2	100
resnet152v2	88.8	75.0	100	87.5	87.5	62.5	100	100	100	12.5	81.8	43.7	87.5	87.5	37.5	100	75.0	100	100	80.6	87.5
resnet50	40.2	25.0	0	100	0	0	37.5	87.5	25.0	0	72.7	37.5	87.5	87.5	100	0	75.0	100	87.5	81.8	87.5
resnet50v2	61.1	50	50	87.5	75.0	25.0	62.5	87.5	25.0	12.5	55.6	25.0	100	0	0	50	100	100	87.5	55.6	62.5
vgg16	27.7	0	37.5	37.5	0	12.5	37.5	50	0	0	34.0	25.0	75.0	12.5	37.5	0	0	50	75.0	34.0	12.5
vgg19	20.8	0	25.0	37.5	12.5	87.5	0	0	0	0	45.4	31.2	0	37.5	25.0	37.5	0	100	75.0	34.0	87.5
xception	91.6	100	100	100	87.5	100	37.5	100	100	0	89.7	100	100	100	87.5	62.5	100	100	100	89.7	62.5

Table 2: Accuracy of Tank (%) trained on grassland backgrounds.

Accuracy (%) for Label : Tank trained on grassland background																					
Models/ Background	desert	desert1	desert2	desert3	desert4	desert5	desert6	desert7	desert8	forest	grass	grass1	grass2	grass3	grass4	grass5	grass6	grass7	grass8	grass9	grass10
DenseNet121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DenseNet169	0	0	0	0	0	0	0	0	0	0	0.19%	1	0	0	0	0	0	0	0	0	0
DenseNet201	0	0	0	0	0	0	0	2.0	0	0	0	0	0	0	0	0	0	0	0	0	0
inceptionresnetv2	2.0	0	2.0	4.6	2.0	0	0	2.0	4.1	0	5.49%	0	10.4	0	0	6.2	0	12.5	18.7	12.5	0
inceptionv3	2.5	0	0	8.3	0	0	0	2.0	8.3	0	3.7	1.0	6.2	0	0	0	2.0	4.0	20.8	8.3	0
mobilenet	6.4	12.5	8.3	0	0	0	12.5	0	20.8	0	13.6	6.2	14.5	0	0	4.1	0	10.4	77.0	29.1	0
mobilenetv2	0.2	0	0	0	0	0	0	0	0	0	0.3	0	0	0	0	0	6.2	4.1	0	0	0
nasnet	6.0	2.0	0	6.2	4.1	2.0	8.3	0	14.5	0	3.8	0	4.1	0	0	4.1	0	4.1	4.0	20.8	0
nasnetmobile	0	0	0	0	0	0	0	0	0	0	0.3	0	0	0	0	0	0	4.0	0	0	0
resnet101	0.2	0	0	0	0	0	0	0	0	0	0.19%	0	0	0	0	0	0	2.0	0	0	0
resnet101v2	2.0	0	2.0	6.2	0	0	2.0	0	4.1	0	0.7	0	0	0	0	0	0	0	0	6.2	0
resnet152	1.3	0	0	4.1	0	0	2.0	0	0	0	0.9	0	2.0	0	0	2.0	0	2.0	2.0	2.0	0
resnet152v2	1.3	0	0	0	0	0	0	2.0	12.5	0	0.7	0	2.0	0	0	0	0	0	0	6.2	0
resnet50	1.1	0	0	6.2	0	0	0	0	0	0	0.5	0	2.0	0	0	0	0	2.0	0	2.0	0
resnet50v2	0.7%	0	0	6.2	0	0	0	0	0	0	0.5	0	0	0	0	0	0	4.1	0	2.0	0
vgg16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
vgg19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
xception	4.8	0	0	10.4	0	0	0	0	14.5	0	7.9	0	39.5	0	0	4.1	14.5	4.1	16.6	8.0	0

Table 3: Accuracy of Tank (%) trained on desert backgrounds.

Accuracy (%) for Label : Tank trained on desert background																					
Models/ Background	desert	desert1	desert2	desert3	desert4	desert5	desert6	desert7	desert8	forest	grass	grass1	grass2	grass3	grass4	grass5	grass6	grass7	grass8	grass9	grass10
DenseNet121	0	0	0	0	0	0	0	0	0	0	1.1	1.0	0	0	0	0	0	10.4	0	0	2.0
DenseNet169	0.4	0	0	0	0	0	0	0	2.0	0	4.1	2.0	10.4	0	2.0	0	8.3	4.1	6.2	10.4	0
DenseNet201	0	0	0	0	0	0	0	0	0	0	0.9	1.0	0	0	0	0	0	6.2	0	2.0	0
inceptionresnetv2	4.3	4.1	4.1	4.1	2.0	2.0	0	12.5	10.4	0	7.0	3.1	12.5	2.0	0	0	4.1	2.0	29.1	20.8	2.0
inceptionv3	3.2	0	0	10.4	0	0	0	6.2	8.3	0	5.4	1.0	10.4	0	0	0	2.0	0	31.4	14.5	0
mobilenet	3.2	2.0	2.8	0	0	0	12.5	2.0	6.2	0	0.3	0	8.3	0	0	0	0	6.2	0	8.3	0
mobilenetv2	0.4	0	2.8	0	0	0	0	0	0	0	0.1	0	0	0	2.0	0	0	4.1	4.1	6.2	0
nasnet	5.5	4.1	2.0	4.1	0	0	2.0	0	20.8	0	1.1	1.0	10.4	0	0	2.5	12.5	6.2	0	18.7	2.0
nasnetmobile	0	0	0	0	2.0	0	0	0	0	0	1.3	0	0	0	0	0	0	2.0	0	2.0	0
resnet101	0.2	0	0	0	0	0	0	0	0	0	1.1	0	0	0	0	0	0	2.0	0	0	0
resnet101v2	0	0	0	0	0	0	0	0	0	0	0.3	0	2.0	0	0	0	4.1	0	4.1	6.2	0
resnet152	1.3	0	2.0	2.0	0	0	0	0	2.0	0	0.9	1.0	2.0	0	2.0	0	0	2.8	2.0	4.1	0
resnet152v2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.0	2.0	0	0	6.2	0
resnet50	1.8	0	0	4.1	0	0	0	4.1	0	0	0.1	0	0	0	0	0	0	0	0	4.1	0
resnet50v2	0.2	0	0	0	0	0	0	0	2.0	0	9.2	0	4.2	0	0	0	4.1	2.0	0	0	0
vgg16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
vgg19	0	0	0	0	0	0	0	0	0	0	1.0	0	0	0	0	0	0	2.0	0	0	0
xception	6.2	0	2.0	12.5	0	0	0	2.0	29.1	0	0.1	0	35.4	0	0	2.0	25.0	4.1	20.8	14.5	0

In this algorithm, many transformations such as flipping, rotation, adding Gaussian noise, etc. are applied the images while optimizing the image to be adversarial. This transformations in the training process of the adversarial image make an adversarial attack more robust against some transformations that might be encountered in real-life scenarios.

However, many of those transformation are limited and are not able to simulate realistic conditions that might affect the adversarial signal when tested in the physical world. For that reason, a better proxy for real life would be working with 3D models, where we can simulate weather conditions, illumination, different poses and so on. Our optimization approach tackles this problem, however, it goes a step further by accounting for the background bias that some models might have. In summary, we created a completed end-to-end optimization process that are robust to changes of camera-views and illumination in different background and scenes.

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[2] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5794–5803, 2018.