# Supplementary Material: A Little Fog for a Large Turn

Harshitha Machiraju, Vineeth N Balasubramanian
Indian Institute of Technology, Hyderabad, India
{ee14btech11011, vineethnb}@iith.ac.in

In this section, we show a few additional results which we could not include in the main paper owing to space constraints. In particular, we show how the proposed method can be used for other datasets, as well as other weather conditions (rain). We also show how our approach to consider task-perceptual similarity can be viewed as a generalization of other methods to construct adversarial perturbations.

## 9. Additional Results

In order to further study the generalizability of our method, we studied the use of our trained model (trained to fool the AutoPilot model on the SullyChen dataset, as in Sec 4) to test on the Udacity dataset [17]. The results are shown in Figure 7. Evidently, the examples show the promise of this approach to generalize to other datasets. While we did not explicitly train or finetune the model for the Udacity dataset in this case, we only believe this will yield more convincing results that can help test steering angle prediction models in such settings.

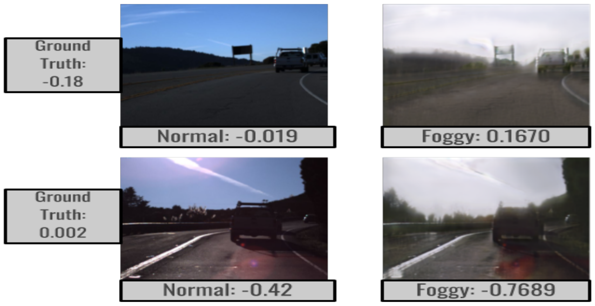

Figure 7: Foggy counterparts of the original images from the Udacity dataset (right) generated using our model pretrained on Sully Chen. The ground truth and predicted steering angles (in radians) in both cases are given alongside.

**Other Weather Conditions:** Our main objective in this work was with fog as the weather condition (which by itself can be valuable) due to the availability of foggy image data, and lack of availability of datasets with sufficient images of other weather conditions in a distinct manner. We, however, studied the usefulness of our approach for other weather conditions by training our model on the SullyChen Dataset with a rainy image video taken from YouTube [1]. The results are shown in Figure 8. Once again, the images show the promise of our method, although the results would be far better with more data with a distinct characterization of the relevant weather condition. The proposed methodology can thus provide a useful framework to test the robustness of autonomous navigation models in different weather conditions.

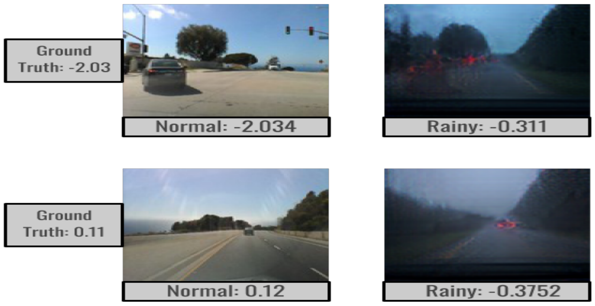

Figure 8: Rainy counterparts of the original images (right) generated using our model trained on Sully Chen with a single rainy video taken from YouTube. The ground truth and predicted steering angles (in radians) in both cases are given alongside.

## 10. A Generalization of Adversarial Attack Methods

As stated in Section 3, the notion of task-perceptual similarity allows us to generate adversarial perturbations that aim to fool a model trained for the corresponding task. We show that it is rather straightforward that existing methods to generate adversarial attacks can be viewed as a special case of this notion, as described below.

**Additive Perturbations:** Additive perturbations are, by far, the most popular of the adversarial attack methods today, where small, carefully crafted perturbations, when added to an images, causes the neural network to misclassify the object. The formal definition of such attacks, given in Eqn 1 of the main paper, was introduced by Szegedy et al in [16]. With time, different methods have been developed for finding $\delta$. Popular methods include: FGSM [6], JSMA [14], ILCM [10], C&W [4], DeepFool [13], and PGD [11]. There have also been approaches which attempt to generalize the perturbation $\delta$ to the entire dataset, i.e., for all inputs of the dataset, a single perturbation is capable of misclassifying all of them. Such methods include: UAP [12], Mopuri et al. [15], Khrulkov et al. [9]. All these methods (more methods can be found in [2]) restrict the adversaries produced to be visually similar to the inputs.

As previously defined, *task-specific perceptual similarity*, the $\phi$ in Eqn 2 (main paper) needs to ensure that the results predicted by the human for both the original and transformed images remains the same. By representing $\phi$ in Eqn 2 as $\phi(\mathbf{x}) = \mathbf{x} + \delta$, this is trivially a specific case of the proposed notion.

**Structural Perturbations:** More recently, adversarial perturbations have shifted from their traditional definition to structural perturbations. These include simple transformations like rotation, translation, and scaling. Some of the popular efforts in this direction are below.

- Engstrom et al. [5] and Kanbak et al. [8] showed that simple rotations are capable of causing adversarial attacks. Further, they also find the required angle which can guarantee that it will be an adversary.
- Azulay et al. [3] show that translation and scaling of images are also capable of causing misclassification.

Contrary to additive perturbations, these transformations are easy to find and produce. By representing $\phi$ in Eqn 2 as a simple rotation, scaling, or translation operator, such perturbations also lie within our notion.

**Textural Perturbations:** Recently, Hendrycks et al. [7] showed that simple texture variations in the background cause the object to be misclassified. They further curated sampled images from the ImageNet dataset, wherein the objects lying in a slightly uncommon background fool the classifier. This work is very much in line with our thought process since adversarially generated fog causes textural changes, which in turn creates a steering deviation. Once again, representing $\phi$ in Eqn 2 as a generative model which is capable of creating objects in new backgrounds, the notion of task-perceptual similarity can be defined as the recognition of the same object in different backgrounds (which humans are capable of, rather easily).

## References

[1] Rain on a car roof-1 hour, 2019. https://www.youtube.com/watch?v=O88fXBx-Qdg. 9

[2] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 10

[3] A. Azulay and Y. Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018. 10

[4] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 10

[5] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017. 10

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 10

[7] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song. Natural adversarial examples, 2019. 10

[8] C. Kanbak, S.-M. Moosavi-Dezfooli, and P. Frossard. Geometric robustness of deep networks: analysis and improvement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4441–4449, 2018. 10

[9] V. Khrulkov and I. Oseledets. Art of singular vectors and universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8562–8570, 2018. 10

[10] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016. 10

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 10

[12] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1765–1773, 2017. 10

[13] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 10

[14] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 10

[15] K. R. Reddy, U. Garg, and V. B. Radhakrishnan. Fast feature fool: A data independent approach to universal adversarial perturbations. *Procedings of the British Machine Vision Conference 2017*, 2017. 10

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 10

[17] Udacity, 2016. https://github.com/udacity/self-driving-car. 9