The Moon can fool your car's autopilot from space in the Autonomous Driving era!

Mechanical Engineering
Carnegie Mellon University

Neel Joshi, Navodit Chandra, Aishwarya Ravi

{ndjoshi, navoditc, aravi2}@andrew.cmu.edu

Carnegie Mellon University

Motivation

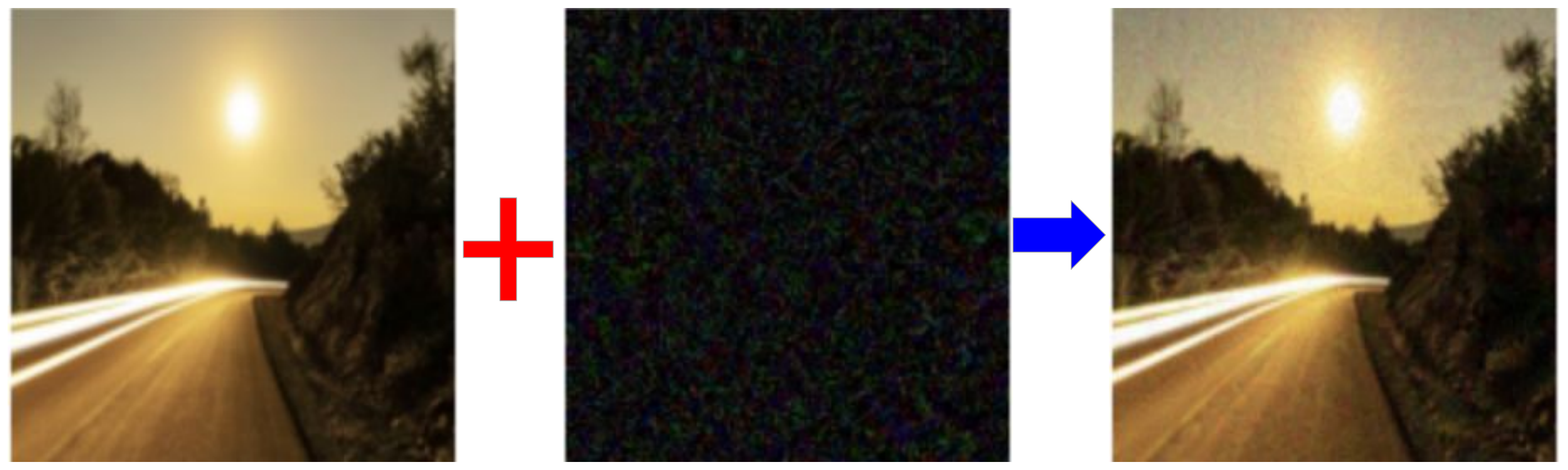
- Tesla FSD mistook the Moon for a traffic light [1], among other such incidents, has attracted thoughts
- We want to demonstrate that such real-life safety-critical scenarios can be avoided

Research Objective

As Safety Engineers, we plan to **reproduce this scenario** **FOR THE FIRST TIME** **using Digital Twin technology and use it as full case study to discuss viable solutions**

Methodology

- The autopilot system of a self-driving car takes decisions on the basis of the images which it captures
- An adversary might exploit this fact to fool the autopilot system by perturbing the images it captures
- We assume that the underlying neural network architecture which the car uses is known apriori and make use of an attack algorithm to perturb the captured images

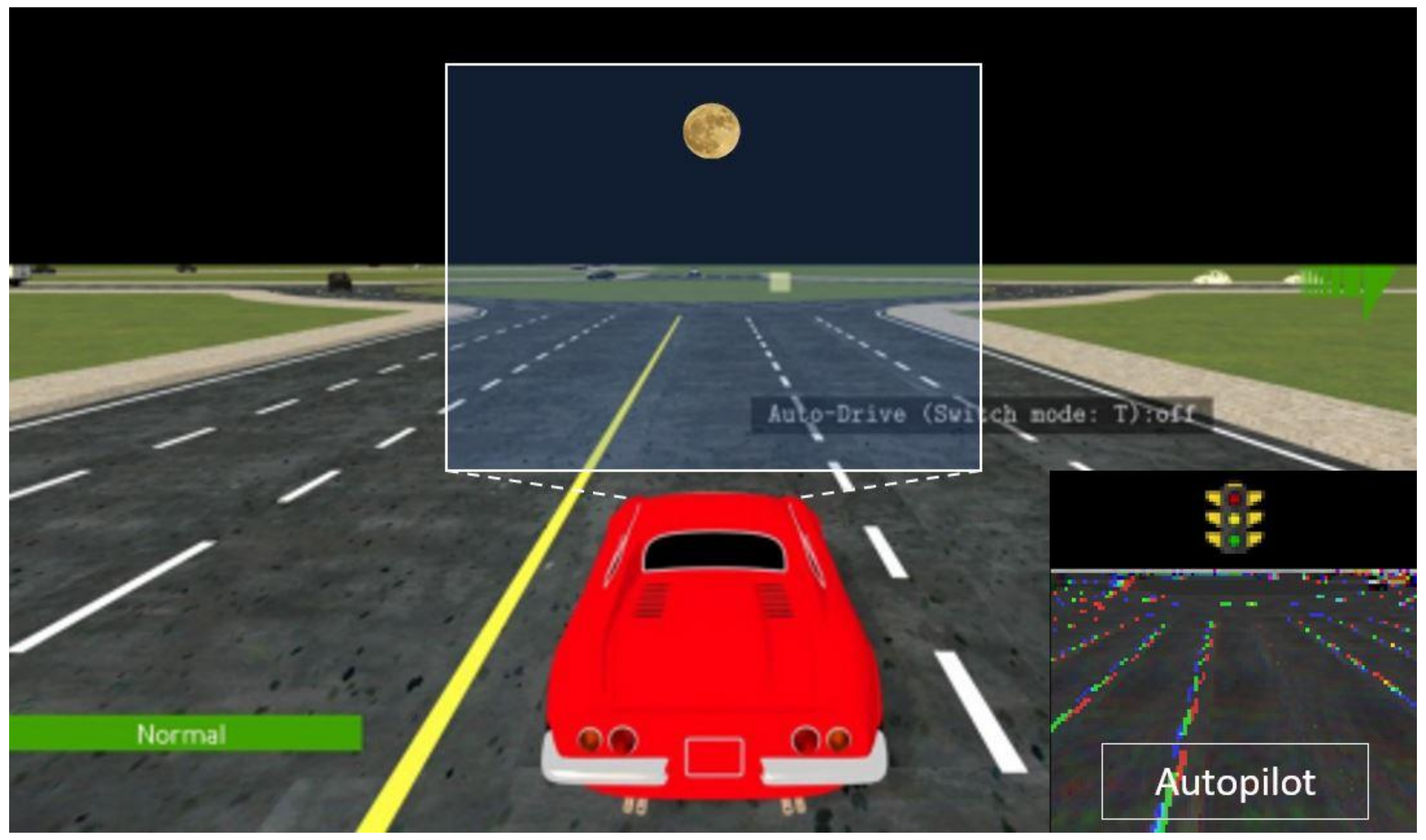


Yellowish tinge in the evening sky

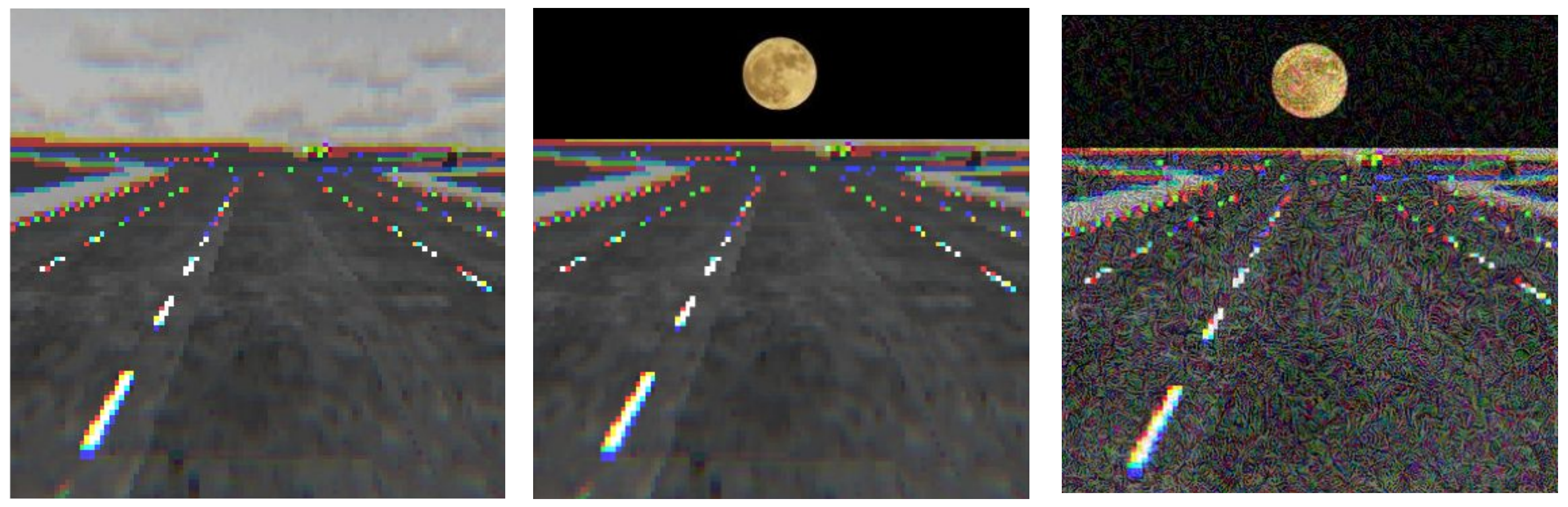
Perturbation generated by the attack algorithm

Yellow traffic light

- For the Digital Twin, we use MetaDrive [2], a light weight, user friendly simulator, for near-to-reality analytics
- Using the source API, we set up a virtual camera on top of the agent (i.e. the car) and extract observations from there



Analytics

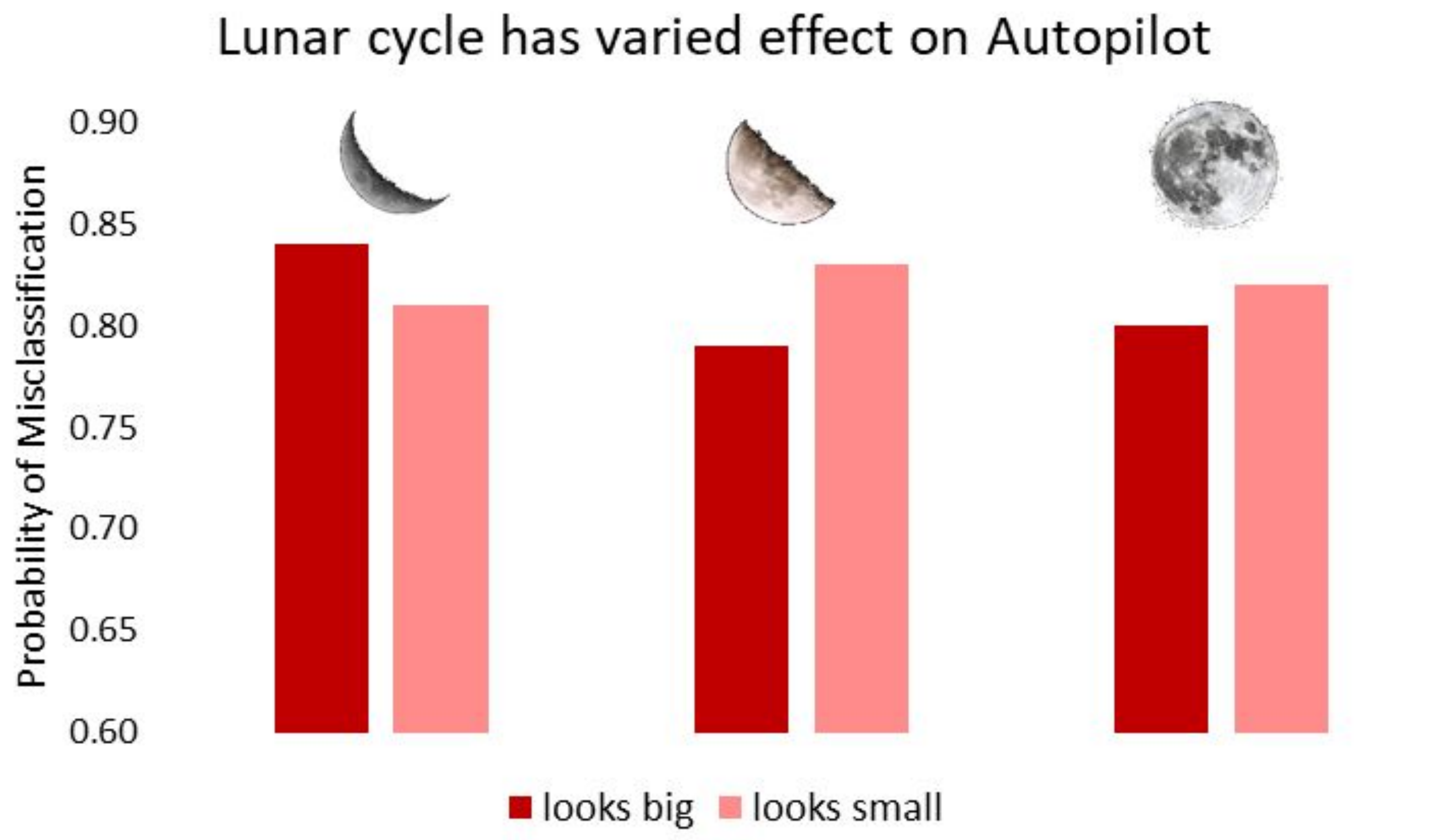


ID = {Airliner}

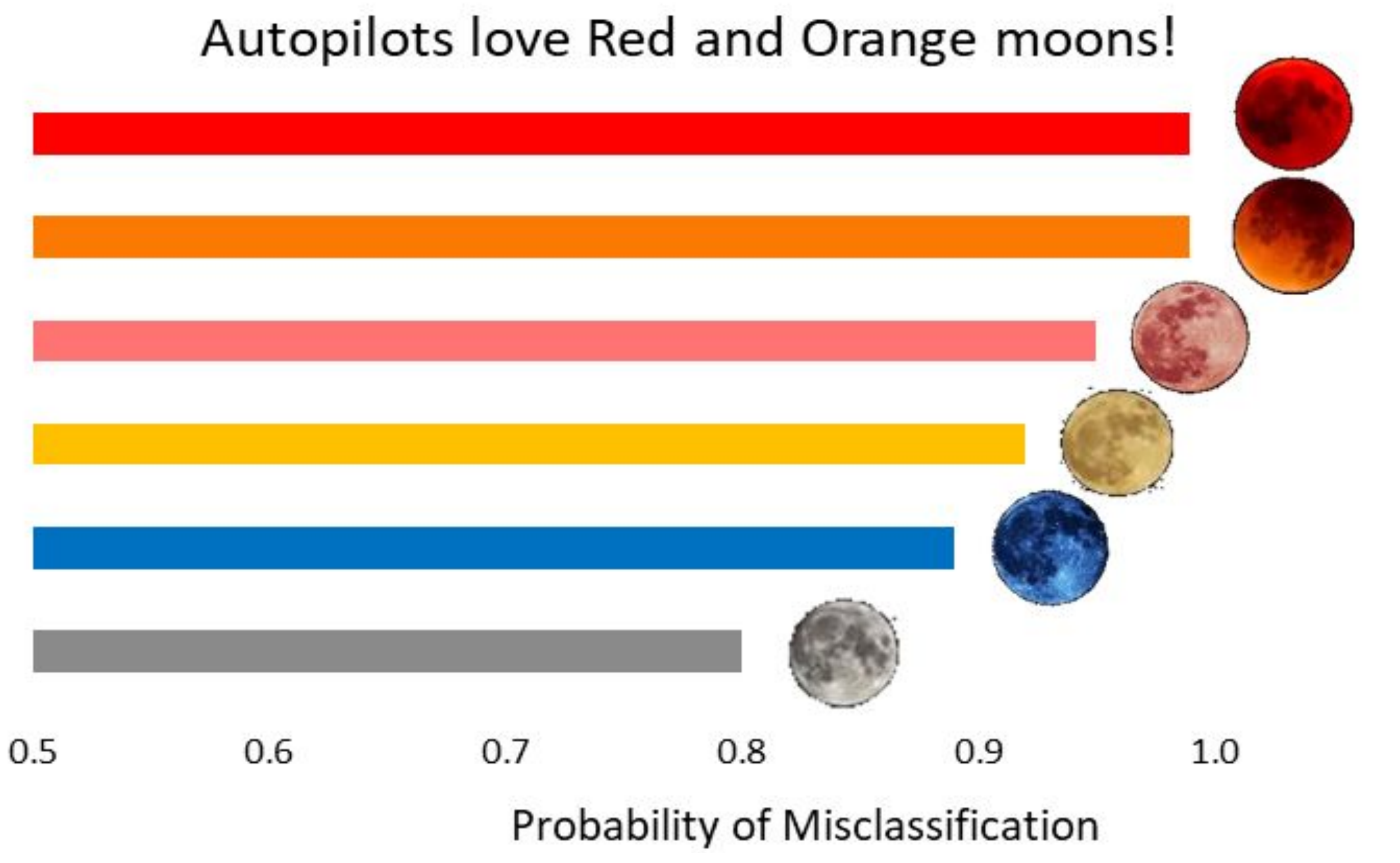
ID = {Balloon}

ID = {TrafficLight}

- The transition of an observation to the adversarial agent is a multifold process
- Along with analysing a real-world safety-critical scenario, we also need to explore possible variations thus giving rise to many more such scenarios



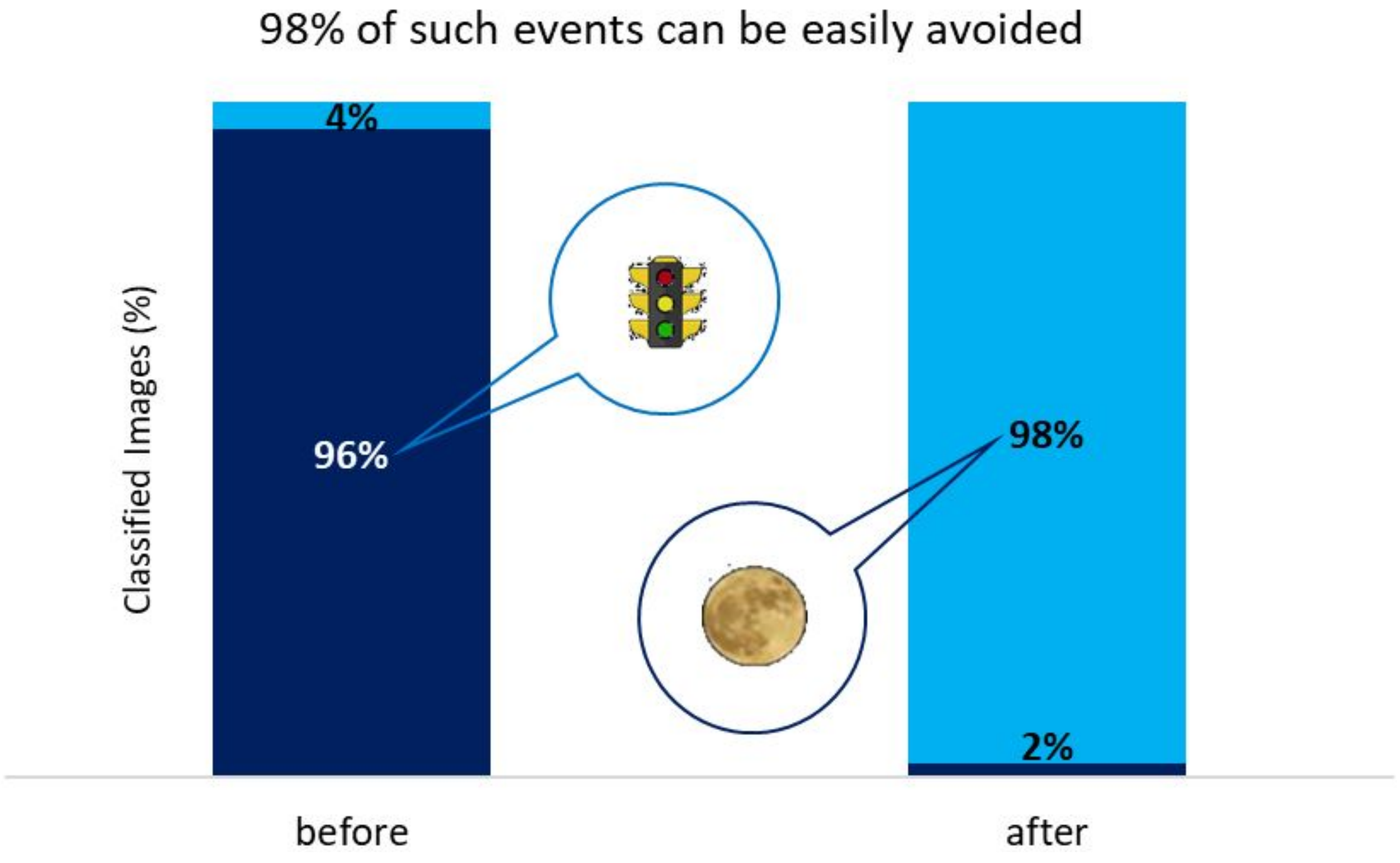
- We tried to understand what effect the colour of the moon can have on the autopilot and predict the period of the year during which this is most likely to happen



Solutions

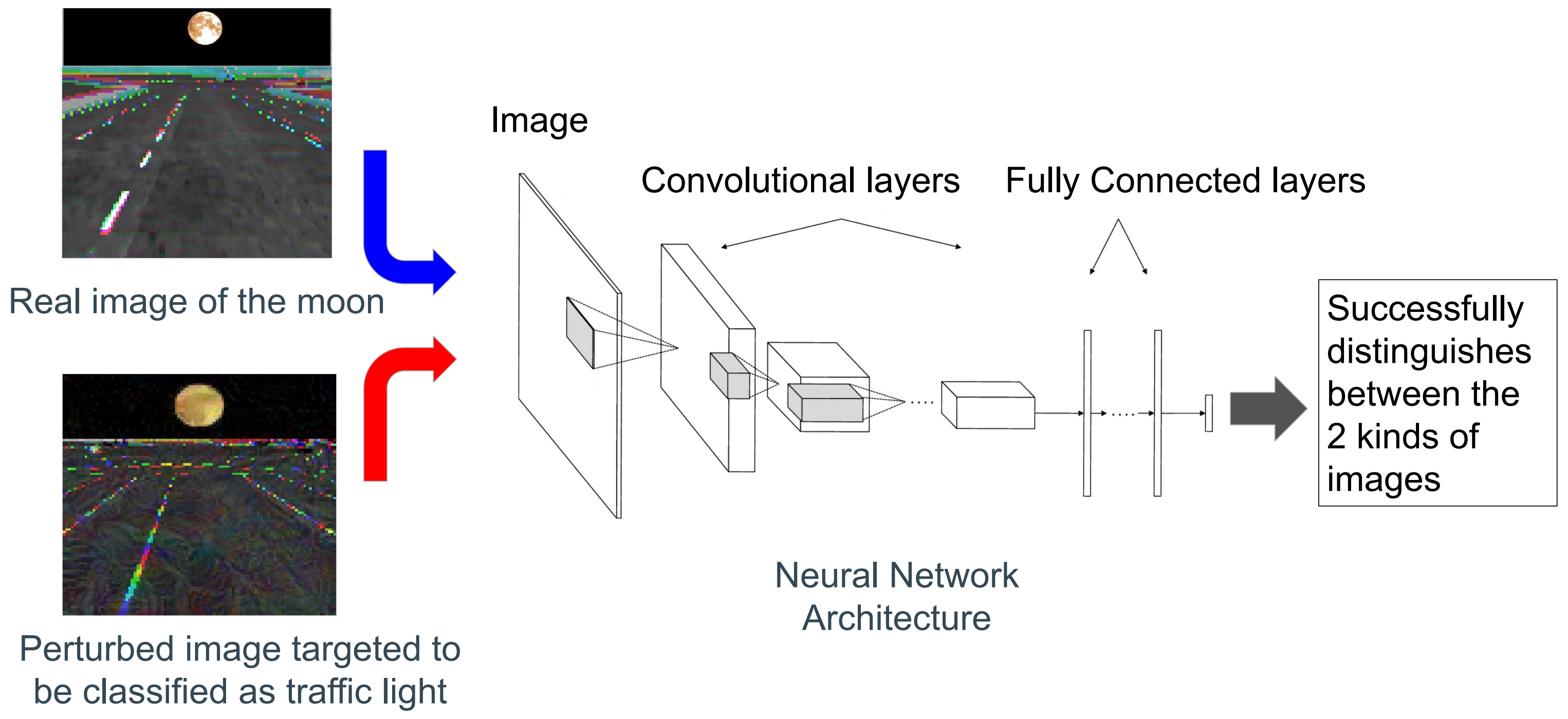
Randomized Padding:

- Input image resized then randomly padded
- Real-time, with low computation, suitable for AVs



Adversarial Training:

- This is the current state of the art against adversarial attacks which can be used as an alternative to random transformations on the images [3].
- It involves training the model with adversarial images such that it becomes robust enough to distinguish between unperturbed and perturbed images.



Real image of the moon
Perturbed image targeted to be classified as traffic light
Image
Convolutional layers
Fully Connected layers
Successfully distinguishes between the 2 kinds of images
Neural Network Architecture

Future Work

- Understand the deeper structure of AV autopilot and find out possible areas for improvement
- Look for more generalizable and robust solutions

References

Jay Ramey, Tesla FSD Mistakes Moon for Yellow Traffic Light (July 23, 2021)

Quanyi Li (2021),"MetaDrive: Composing Diverse Driving Scenarios for Generalizable Reinforcement Learning.". In: Arxiv, Machine Learning.

Oscar Knagg, Know your enemy (January 6, 2019)

Hashemi, M. (2019). Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. *Journal of Big Data*, *6*(1), 1-13.