Putting Image Manipulations in Context: Robustness Testing for Safe Perception

Trenton Tabor, Senior Robotics Engineer
National Robotics Engineering Center

Zachary Pezzementi, **Trenton Tabor**, Samuel Yim, Jonathan K. Chang, Bill Drozd, David Guttendorf, Michael Wagner and Philip Koopman

"Robustness Inside-Out Testing (RIOT)." NAVAIR Public Release 2018-165. Distribution Statement A – "Approved for public release; distribution is unlimited"

Acknowledgements

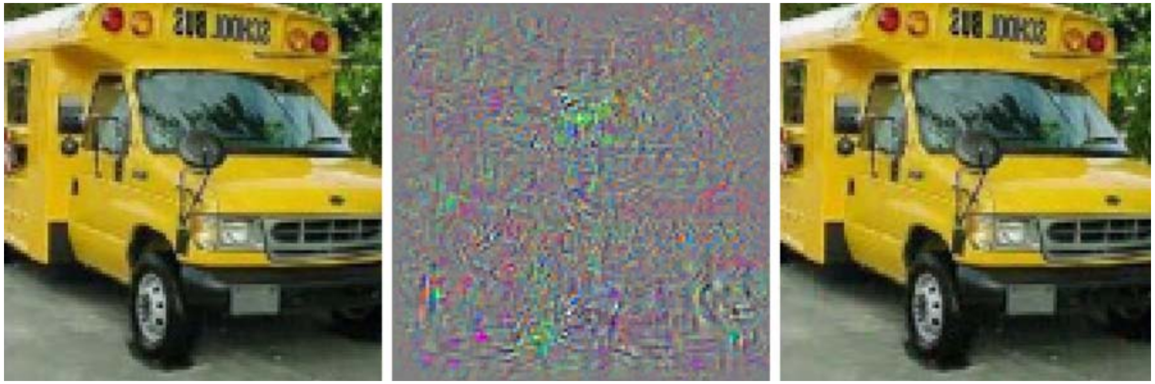
This project was funded by the Test Resource Management Center (TRMC) Test and Evaluation / Science & Technology (T&E/S&T) Program through the U.S. Army Program Executive Office for Simulation, Training and Instrumentation (PEO STRI) under Contract No. W900KK-16-C-0006, "Robustness Inside-Out Testing (RIOT)." NAVAIR Public Release 2018-165. Distribution Statement A – "Approved for public release; distribution is unlimited".

We also thank other authors for making code public to enable this work.

Adversarial Perturbations

We already know modern object detectors' behavior can change drastically from small changes to an image.

Those behavior changes can be critical to safety.



Detected as: "Bus" | Adversarial Noise | Detected as: "Ostrich"

Szegedy et al. "Intriguing properties of neural networks". arXiv 1312.6199

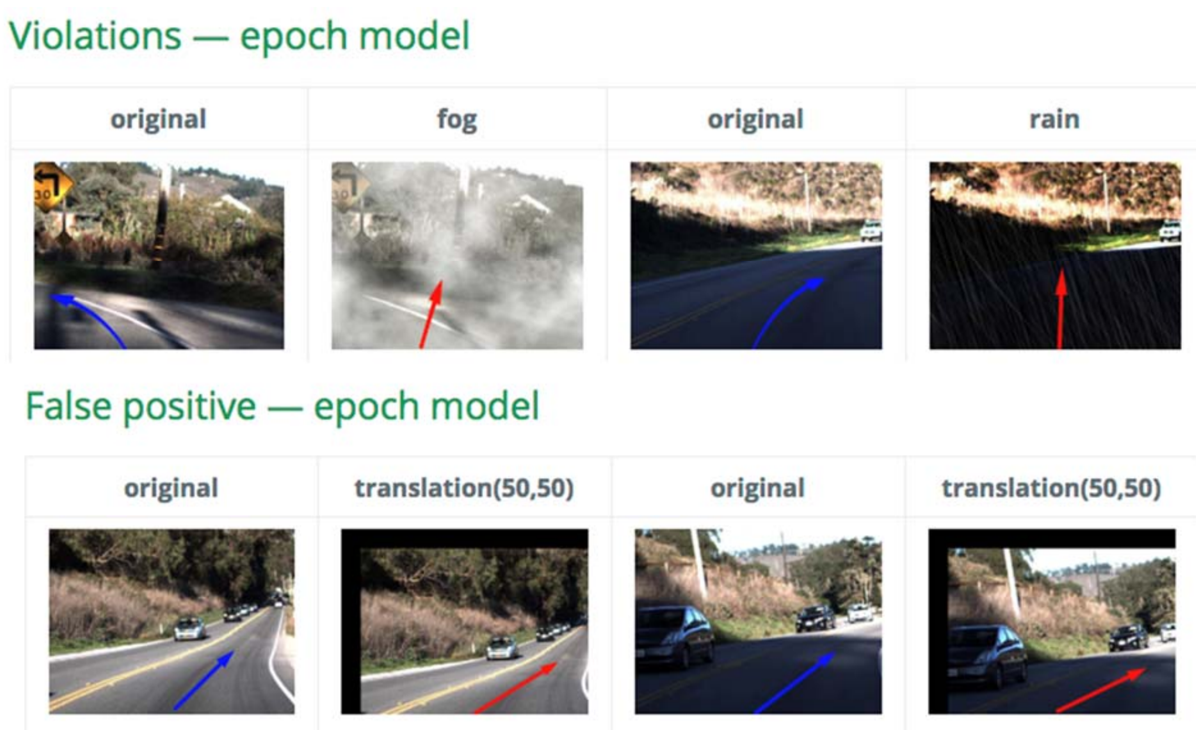
But what if we are not worried about adversarial inputs?

Related Work - DeepTest

Apply mutations to images and evaluate effect on system learned to produce steering angles

Our approach's differences:

- Physically realistic mutators
  - Better match what real systems may encounter
  - Ensure ideal output should not change
- Large-scale evaluation
  - Dataset size
  - # Mutators
  - # Algorithms
- Evaluate person detection



Y Tian et al. "Deeptest: Automated testing of deep-neural-network-driven autonomous cars." ICSE, 2018.

Major Take-Aways

Small image changes can have catastrophic effects on safety critical perception.

In fact, common image degradations can often cause such failures for systems running over long periods.

We demonstrate this on many state-of-art fieldable systems within a framework for evaluating robustness in adverse conditions.

Previous Work – Person Detection

Person detection is one sample safety-critical application now dominated by deep-learning-based approaches.

Experiments in this work use NREC Agricultural Person Detection Dataset, largest public dataset for off-road person detection

T. Tabor et al. "People in the weeds: Pedestrian detection goes off-road." In *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR).*

Z. Pezzementi et al. "Comparing apples and oranges: Off-road pedestrian detection on the National Robotics Engineering Center agricultural person-detection dataset." *Journal of Field Robotics*, 2018.

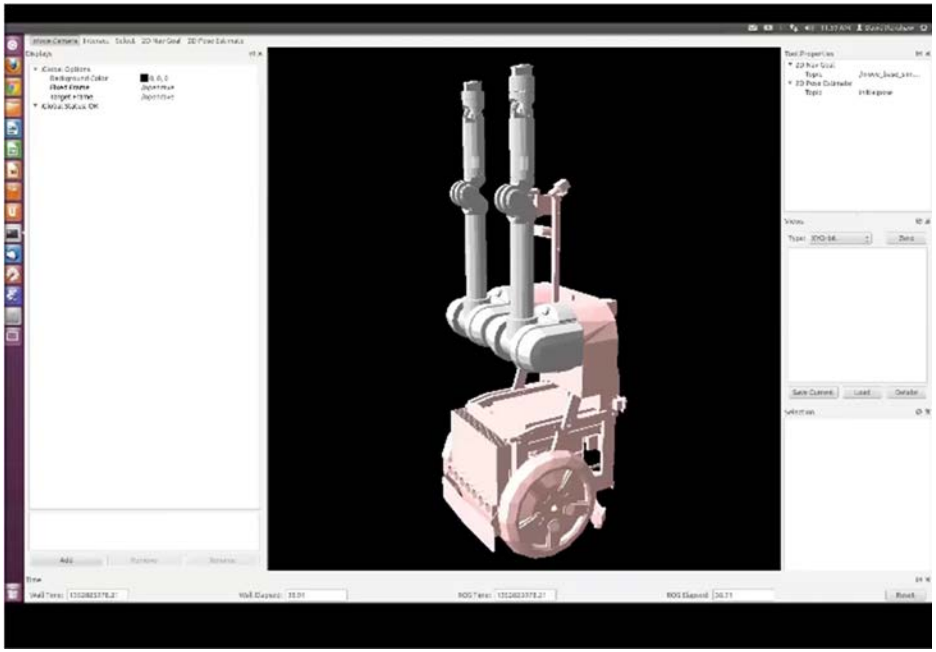
Philosophy of Approach – Robustness Testing for Perception

Chose to partner with the robustness testing group at our University

Robustness Testing is the process of generating many queries of a system and requires knowledge of what wrong behavior is for these inputs.

This technique has found real, dangerous bugs

- Previously tested 17 robotic systems over several years
- An example, shown, is finding a planner ignoring constraints, leading to erratic behavior



C. Hutchison et al. “Robustness testing of autonomy software.” *International Conference on Software Engineering*, 2018.

Applying to perception…

- Perception systems have such high dimensional inputs to make pure generation of new inputs impractical.
- We propose instead using physically grounded mutations of previously labeled data to create exceptional inputs.

Details of Approach

From Philosophy

We propose using **physically grounded mutations** of previously labeled data to create exceptional inputs.

- **We built a list of degradations that occur in outdoor imaging**
- **We implemented parameterized mutators modeling some of these**
- **We used these mutated datasets to estimate robustness of safety critical machine learning systems**

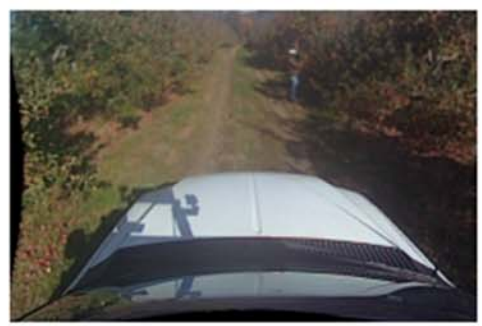
Mutators Used

(a) Original Image, $\mathbf{I}$ (b) Gaussian Blur, $\sigma = 2$ (c) Brightness Shift, $b = 0.5$ (d) Alpha Blend, $\alpha = 0.5$

(e) JPEG Compression, $q = 40\%$ (f) Salt and Pepper, $s = 10\%$ (g) Channel Drop-Out, $Y(Cb)Cr$ (h) Additive Gaussian Noise, $\sigma_w = 15.0$, $\sigma_u = 0.5$, $\gamma = 0.5$

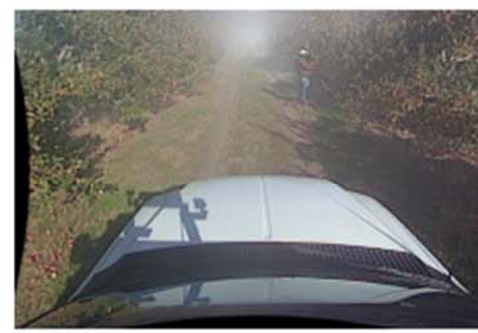
All are literature backed degradations

See paper for details and model references



$u_f = 1\text{m}, \kappa = 2$

Defocus



$u_V = 97.8\text{m}$

Haze

Contextual Mutators

Contextual Mutators

We implemented two mutators that depend on the geometric structure of the scene, defocus and haze:

- Estimate scene geometry in each frame from stereo video using scene flow[1] and bilateral filtering[2]
- Mutate images based on estimated depth to each pixel:
  - Defocus – Based on depth-dependent blurring
  - Haze – Based on depth-dependent alpha-blending

[1] Vogel et al. "3D Scene Flow Estimation with a Piecewise Rigid Scene Model". IJCV 2015.
[2] Barron and Poole. "The Fast Bilateral Solver". ECCV 2016.

Mesh of typical reconstruction

Typical defocus mutation

Detectors Tested

Many state-of-the-art object detectors from some of the most popular deep network frameworks

SUT	Base Library	Reference(s)
MS-CNN	Caffe	[11]
SSD w/ MobileNets	TensorFlow	[12]–[14]
SSD w/ Inception	TensorFlow	[12], [13], [15]
R-FCN w/ ResNet-101	TensorFlow	[12], [16], [17]
Faster R-CNN w/ ResNet-101	TensorFlow	[12], [17], [18]
Faster R-CNN w/ Inception ResNet v2	TensorFlow	[12], [18], [19]
Deformable R-FCN	MXNet	[16], [20]
Deformable Faster R-CNN	MXNet	[18], [20]

Striking Results

Small changes to images, even physically realistic changes, can cause a catastrophic change in classifier performance

Ground Truth Label

Raw Detection Strength

Required False Positive Rate

Original

Mutated

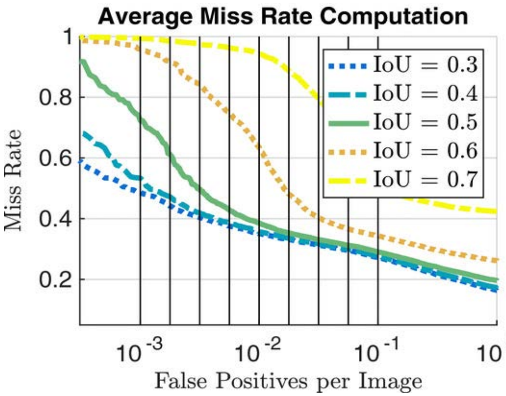
MSCNN detections on original images and under moderate blur

How We Evaluate Detection Performance

Detections and ground truth are bounding boxes around people.

(a) Detections failing $T = 0.5$ (b) Detections failing $T = 0.3$

We consider multiple Intersection-over-Union (IoU) thresholds for whether to consider a detection "correct".



Then we average area under the ROC curve for each to get a single number for overall detection performance.

MS-CNN	SSD w/ MobileNets	SSD w/ Inception	Faster R-CNN w/ Resnet 101	R-FCN w/ Resnet 101	Faster R-CNN w/ Inception Resnet	Deformable R-FCN	Deformable Faster R-CNN
0.60	0.29	0.22	0.64	0.64	0.71	0.71	0.73

Z. Pezzementi et al. "Comparing apples and oranges: Off-road pedestrian detection on the National Robotics Engineering Center agricultural person-detection dataset." *Journal of Field Robotics*, 2018.

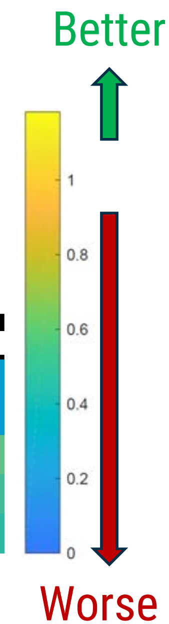
Sample Results: Channel Dropout

Channel dropout is devastating to most detectors

Original Image

Drop Channel Cb (YCbCr)

Mutator & Parameters	MS-CNN	SSD w/ MobileNets	SSD w/ Inception	Faster R-CNN w/ Resnet 101	R-FCN w/ Resnet 101	Faster R-CNN w/ Inception Resnet	Deformable R-FCN	Deformable Faster R-CNN
Baseline	**0.60**	**0.29**	**0.22**	**0.64**	**0.64**	**0.71**	**0.71**	**0.73**
Drop Channel Cb (YCbCr)	0.36	0.01	0.00	0.40	0.09	0.41	0.16	0.11
Drop Channel Cr (YCbCr)	0.30	0.00	0.00	0.33	0.04	0.49	0.13	0.10
Drop Channel R (RGB)	0.64	0.07	0.01	0.51	0.34	0.56	0.34	0.37
Drop Channel G (RGB)	0.49	0.03	0.00	0.45	0.23	0.60	0.28	0.32
Drop Channel B (RGB)	0.40	0.03	0.03	0.39	0.23	0.58	0.29	0.29



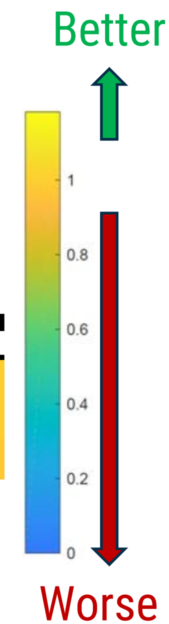
Sample Results: Haze

There is variation in detector robustness to haze

Original Image

Haze w/ 97.8m Visibility (β 0.04)

Mutator & Parameters	MS-CNN	SSD w/ MobileNets	SSD w/ Inception	Faster R-CNN w/ Resnet 101	R-FCN w/ Resnet 101	Faster R-CNN w/ Inception Resnet	Deformable R-FCN	Deformable Faster R-CNN
Baseline	**0.60**	**0.29**	**0.22**	**0.64**	**0.64**	**0.71**	**0.71**	**0.73**
Haze (u_V 978.0 m (β 0.004))	0.56	0.29	0.22	0.64	0.64	0.69	0.63	0.73
Haze (u_V 326.0 m (β 0.012))	0.50	0.28	0.21	0.64	0.65	0.67	0.63	0.73
Haze (u_V 97.8 m (β 0.04))	0.36	0.19	0.14	0.61	0.60	0.61	0.61	0.71



Full Results

Evaluated full combination of mutators and detectors

Allows analysis of general robustness characteristics of each detector

Sometimes would change choice of best detector, depending on importance of adverse conditions to you

Mutator & Parameters	MS-CNN	SSD w/ MobileNets	SSD w/ Inception	Faster R-CNN w/ Resnet 101	R-FCN w/ Resnet 101	Faster R-CNN w/ Inception Resnet	Deformable R-FCN	Deformable Faster R-CNN
Baseline	**0.60**	**0.29**	**0.22**	**0.64**	**0.64**	**0.71**	**0.71**	**0.73**
Defocus (u_f 10.0; κ 2.0)	0.59	0.29	0.22	0.64	0.63	0.71	0.63	0.73
Defocus (u_f 5.0; κ 2.0)	0.59	0.29	0.22	0.64	0.64	0.71	0.63	0.73
Defocus (u_f 2.0; κ 2.0)	0.52	0.29	0.24	0.63	0.63	0.68	0.62	0.74
Defocus (u_f 1.0; κ 2.0)	0.38	0.20	0.21	0.54	0.53	0.57	0.50	0.69
Defocus (u_f 10; κ 2.8)	0.59	0.29	0.22	0.64	0.63	0.71	0.63	0.73
Defocus (u_f 5; κ 2.8)	0.59	0.29	0.23	0.64	0.64	0.71	0.63	0.73
Defocus (u_f 2; κ 2.8)	0.47	0.26	0.23	0.59	0.59	0.65	0.58	0.73
Defocus (u_f 1.0; κ 2.8)	0.27	0.14	0.17	0.47	0.44	0.44	0.40	0.58
Defocus (u_f 10.0; κ 3.6)	0.59	0.29	0.22	0.64	0.63	0.71	0.63	0.73
Defocus (u_f 5.0; κ 3.6)	0.57	0.29	0.23	0.64	0.63	0.70	0.62	0.73
Defocus (u_f 2.0; κ 3.6)	0.43	0.24	0.22	0.55	0.56	0.60	0.53	0.70
Defocus (u_f 1.0; κ 3.6)	0.19	0.11	0.13	0.42	0.38	0.36	0.34	0.51
Gaussian Blur (σ 0.5)	0.56	0.29	0.23	0.64	0.64	0.70	0.63	0.74
Gaussian Blur (σ 1.0)	0.48	0.27	0.24	0.61	0.61	0.67	0.60	0.74
Gaussian Blur (σ 1.5)	0.41	0.22	0.22	0.56	0.56	0.61	0.54	0.71
Gaussian Blur (σ 2.0)	0.33	0.17	0.19	0.51	0.49	0.53	0.47	0.65
Gaussian Blur (σ 2.5)	0.25	0.13	0.16	0.47	0.44	0.45	0.41	0.59
Gaussian Blur (σ 3.0)	0.19	0.10	0.14	0.43	0.40	0.37	0.35	0.53
Haze (u_V 978.0 m (β 0.004))	0.56	0.29	0.22	0.64	0.64	0.69	0.63	0.73
Haze (u_V 326.0 m (β 0.012))	0.50	0.28	0.21	0.64	0.65	0.67	0.63	0.73
Haze (u_V 97.8 m (β 0.04))	0.36	0.19	0.14	0.61	0.60	0.61	0.61	0.71
Alpha Blend (α 0.1)	0.53	0.29	0.21	0.64	0.64	0.69	0.63	0.73
Alpha Blend (α 0.25)	0.38	0.24	0.18	0.64	0.62	0.66	0.63	0.73
Alpha Blend (α 0.5)	0.22	0.05	0.09	0.63	0.55	0.63	0.63	0.72
Alpha Blend (α 0.75)	0.21	0.00	0.00	0.54	0.28	0.55	0.59	0.67
JPEG Compression (q 40)	0.56	0.27	0.21	0.62	0.61	0.68	0.61	0.71
JPEG Compression (q 20)	0.51	0.25	0.19	0.57	0.57	0.64	0.58	0.68
JPEG Compression (q 10)	0.39	0.19	0.15	0.47	0.46	0.51	0.49	0.58
Brightness (b 2.00)	0.61	0.14	0.09	0.51	0.59	0.60	0.59	0.66
Brightness (b 1.33)	0.63	0.25	0.16	0.60	0.64	0.66	0.63	0.72
Brightness (b 1.14)	0.61	0.27	0.19	0.62	0.64	0.69	0.63	0.73
Brightness (b 0.88)	0.57	0.30	0.25	0.65	0.63	0.72	0.62	0.73
Brightness (b 0.75)	0.55	0.30	0.26	0.64	0.62	0.73	0.62	0.72
Brightness (b 0.50)	0.56	0.24	0.23	0.61	0.58	0.73	0.60	0.71
Salt and Pepper (1% of pixels)	0.58	0.27	0.20	0.60	0.61	0.66	0.61	0.70
Salt and Pepper (2% of pixels)	0.55	0.25	0.18	0.57	0.59	0.63	0.60	0.68
Salt and Pepper (5% of pixels)	0.50	0.21	0.14	0.51	0.54	0.58	0.55	0.61
Drop Channel Cb (YCbCr)	0.36	0.01	0.00	0.40	0.09	0.41	0.16	0.11
Drop Channel Cr (YCbCr)	0.30	0.00	0.00	0.33	0.04	0.49	0.13	0.10
Drop Channel R (RGB)	0.64	0.07	0.01	0.51	0.34	0.56	0.34	0.37
Drop Channel G (RGB)	0.49	0.03	0.00	0.45	0.23	0.60	0.28	0.32
Drop Channel B (RGB)	0.40	0.03	0.03	0.39	0.23	0.58	0.29	0.29
Additive (ζ_w 5.0; ζ_u 0.5; ψ 0.5)	0.60	0.28	0.21	0.63	0.62	0.69	0.62	0.71
Additive (ζ_w 5.0; ζ_u 0.5; ψ 0.7)	0.60	0.27	0.19	0.61	0.60	0.66	0.60	0.68
Additive (ζ_w 5.0; ζ_u 1.5; ψ 0.5)	0.60	0.26	0.19	0.61	0.59	0.65	0.59	0.66
Additive (ζ_w 15.0; ζ_u 0.5; ψ 0.5)	0.59	0.25	0.18	0.60	0.58	0.65	0.59	0.66
Additive (ζ_w 5.0; ζ_u 2.5; ψ 0.5)	0.59	0.21	0.15	0.56	0.54	0.60	0.55	0.60

Predicting Contextual Mutators

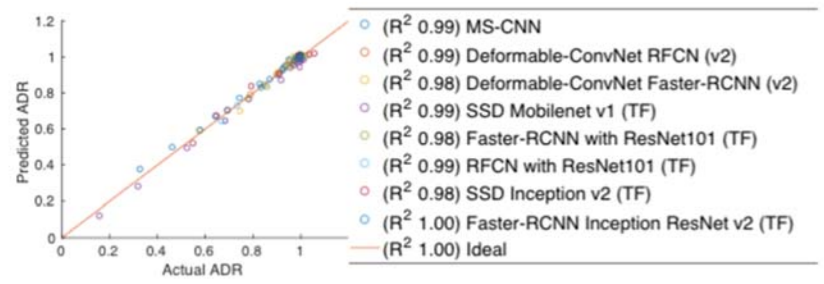
For each contextual mutator, we have a simple equivalent that does not require geometric context:

Simple	Contextual
Gaussian Blur	Defocus
Alpha Blend	Haze

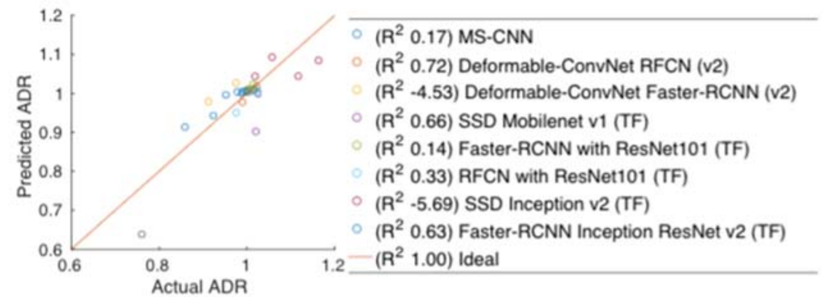
Can performance under contextual mutators be predicted from simple mutators?

Works well for predicting Defocus from Gaussian Blur

Works poorly for predicting Haze from Alpha Blend



(a) Defocus as predicted by blur



(b) Haze as predicted by alpha blend

Threats to Validity (And Future Work)

- **Generalization outside our chosen detection algorithms**
  - New systems are developed all the time, and each has a configuration space
- **Generalization across performance metrics**
  - We chose detection accuracy, but there are many other options
- **Generalization outside our dataset**
  - Focused on an off-road dataset, there are many domains where autonomous vehicles are applied
- **Generalization outside our chosen mutators**
  - Future mutators may have radically different effects on detectors

Major Take-Aways

Small image changes can have catastrophic effects on safety critical perception.

In fact, common image degradations can often cause such failures for systems running over long periods.

We demonstrate this on many state-of-art fieldable systems within a framework for evaluating robustness in adverse conditions.