ICCV 2023 Tutorial

The Many Faces of Reliability of Deep Learning for Real-World Deployment

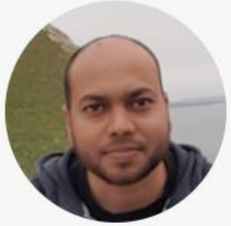
Tuesday, October 3rd 2023, 08:30 - 13:00
Room S05
Paris, France

Andrei Bursuc
valeo.ai

Tuan-Hung Vu
valeo.ai

Sharon Yixuan Li
UW-Madison

Dengxin Dai
Huawei



Puneet Dokania
U. Oxford, Five AI

Patrick Pérez
valeo.ai

ICCV 2023 Tutorial

🎭 The Many Faces of Reliability of Deep Learning for Real-World Deployment 🌍

Tuesday, October 3rd 2023, 08:30 - 13:00
Room S05
Paris, France

Schedule

08:30 - 08:50 **Setting the stage: from academic benchmarks to real-world situations** by **Patrick**

08:50 - 09:25 **Uncertainty estimation and next generation ensembles** by **Andrei**

09:25 - 10:20 **Calibration of Deep Neural Networks** by **Puneet**

10:20 - 10:40 **Break**

10:40 - 11:35 **Out-of-distribution detection** by **Sharon**

11:35 - 12:30 **Robustness and generalization under distribution shift** by **Dengxin** and **Tuan-Hung**

12:30 - 12:45 **Performance monitoring** by **Andrei**

12:45 - 13:00 **Closing remarks + Q&A** by **All**


Café Journal
Bistro Ristorante Italiano
OFF LICENCE
cash
Holborn
Liverpool Street 8
Bow Church
Stratford City 388


96 Porte des Lilas

1.36M

deaths due to vehicle crashes each year

42,915

deaths in the U.S. in 2021 and 2.5 million injuries

$836B

in harm from loss of life and injury each year

50M

Injuries worldwide due to vehicle crashes each year

79%

of seniors age 65 and older live in car-dependent communities.

12M

people 40 years and over in the United States have vision impairment.

Source: https://waymo.com/safety/

20+

Million Autonomously Driven Miles

25+

Cities

10

U.S. States

15+

Billion Simulated Miles

Autonomous Software Testing

Like our hardware, our autonomous driving software is guided by our *Safety by Design* philosophy. We constantly and rigorously test the individual components of the software—including perception, behavior prediction, and planner—as well as the software as a whole.

Our technology is constantly learning and improving. Each change of our software undergoes a rigorous release process and is tested through a combination of simulation testing, closed course testing, and driving on public roadways:

Simulation Testing

In simulation, we rigorously test any changes or updates to our software before they're deployed in our fleet. We identify the most challenging situations our vehicles have encountered on public roads, and turn them into virtual scenarios for our autonomous driving software to practice in simulation. We also review data from crash databases and naturalistic driving studies to identify other possible collision scenarios and develop tests accordingly.

Closed-Course Testing

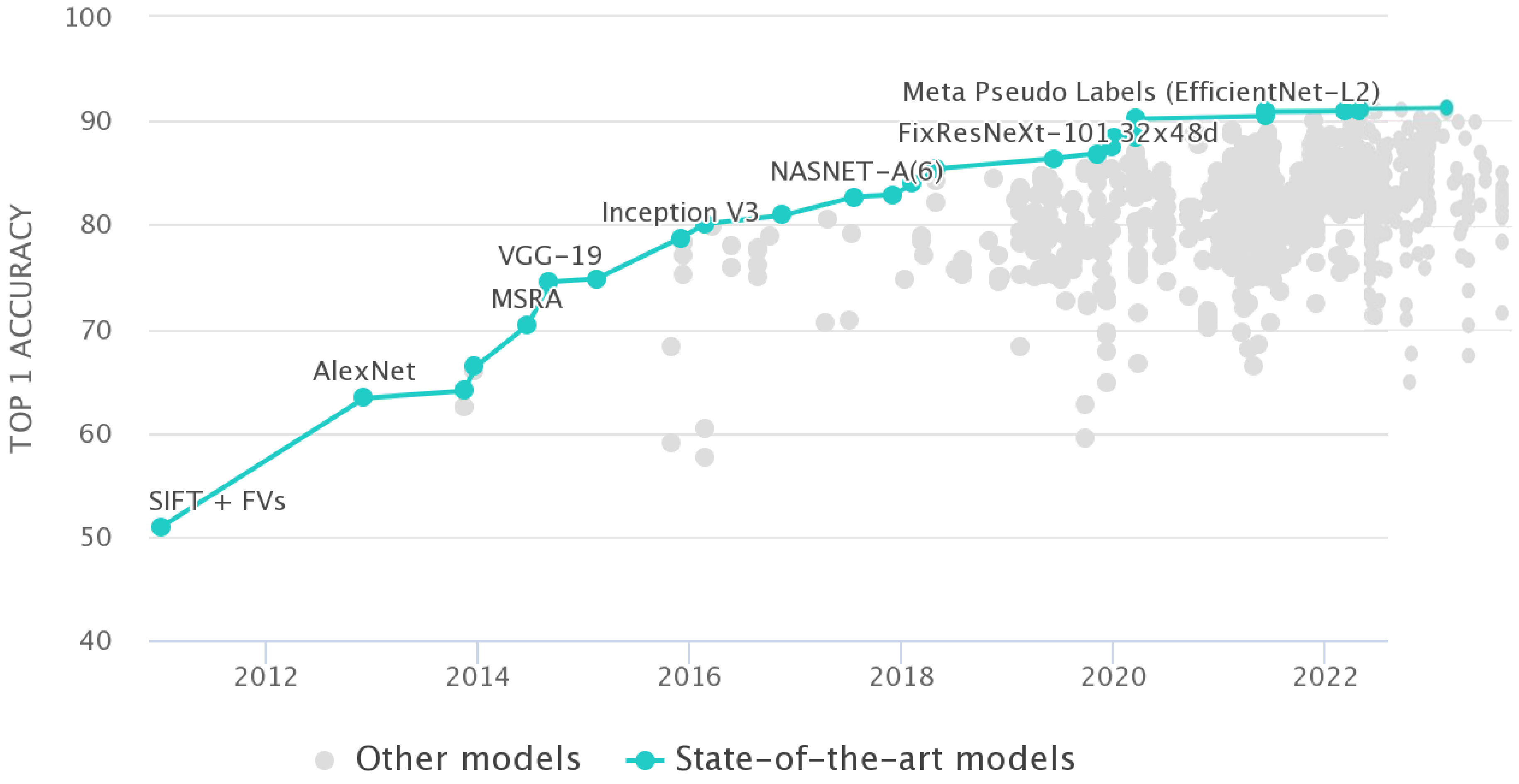
New software is pushed to a few vehicles first so that our most experienced drivers can test the new software, typically starting on our private test track. We can use different releases of software for different vehicles so that we can test new or specific features within different operational design domains.

Real-World Driving

Once we confirm that our software is working as intended, we begin introducing the new software to our vehicles on public roads. We start small and then gradually push the software update to our entire fleet after we've gained greater confidence in its performance. The more miles we travel on public roads, the more opportunities to monitor and assess the performance of software.

Source: https://waymo.com/safety/

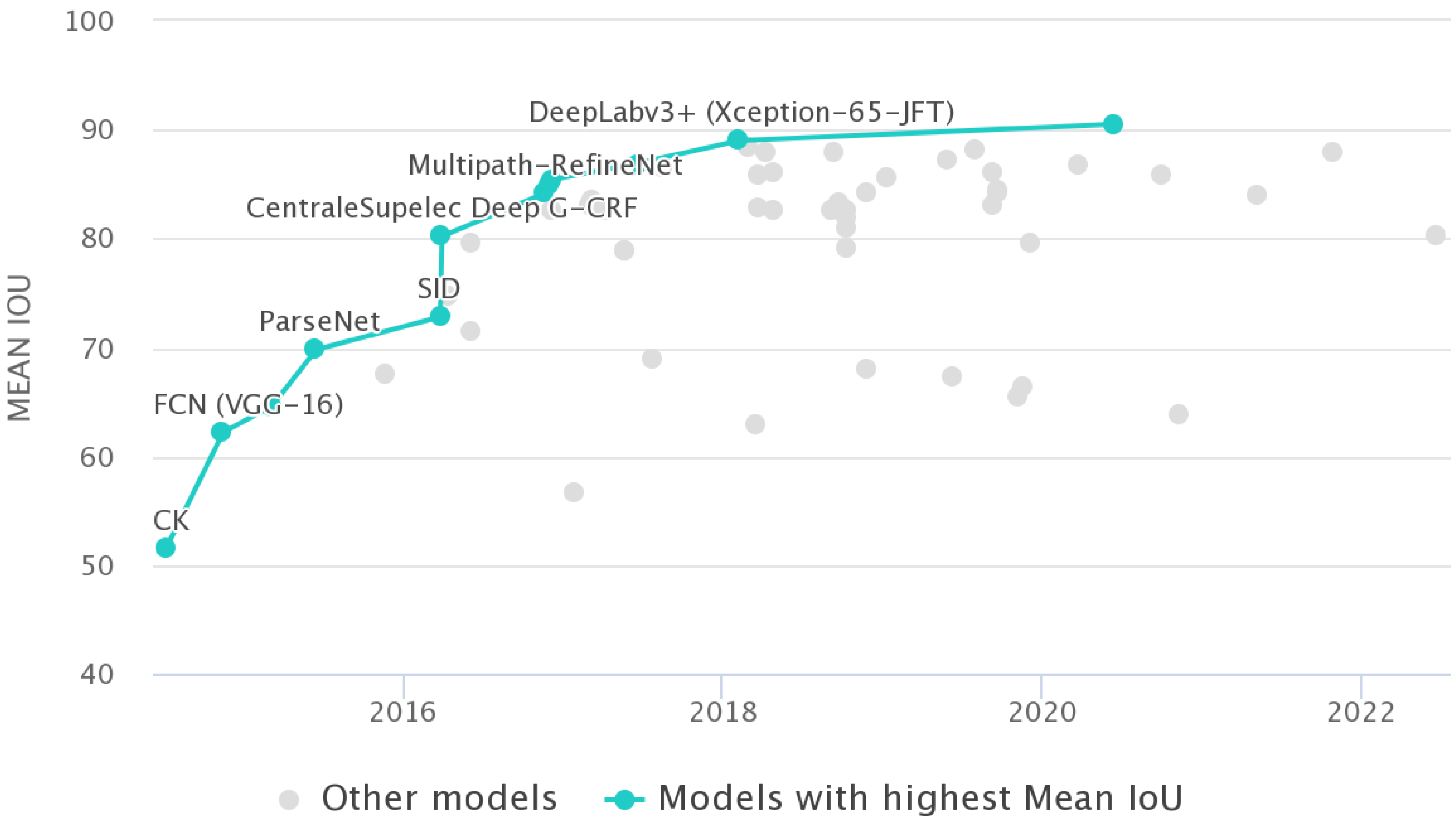
A decade of mad progress



ImageNet-1K image classification

https://paperswithcode.com/sota/image-classification-on-imagenet

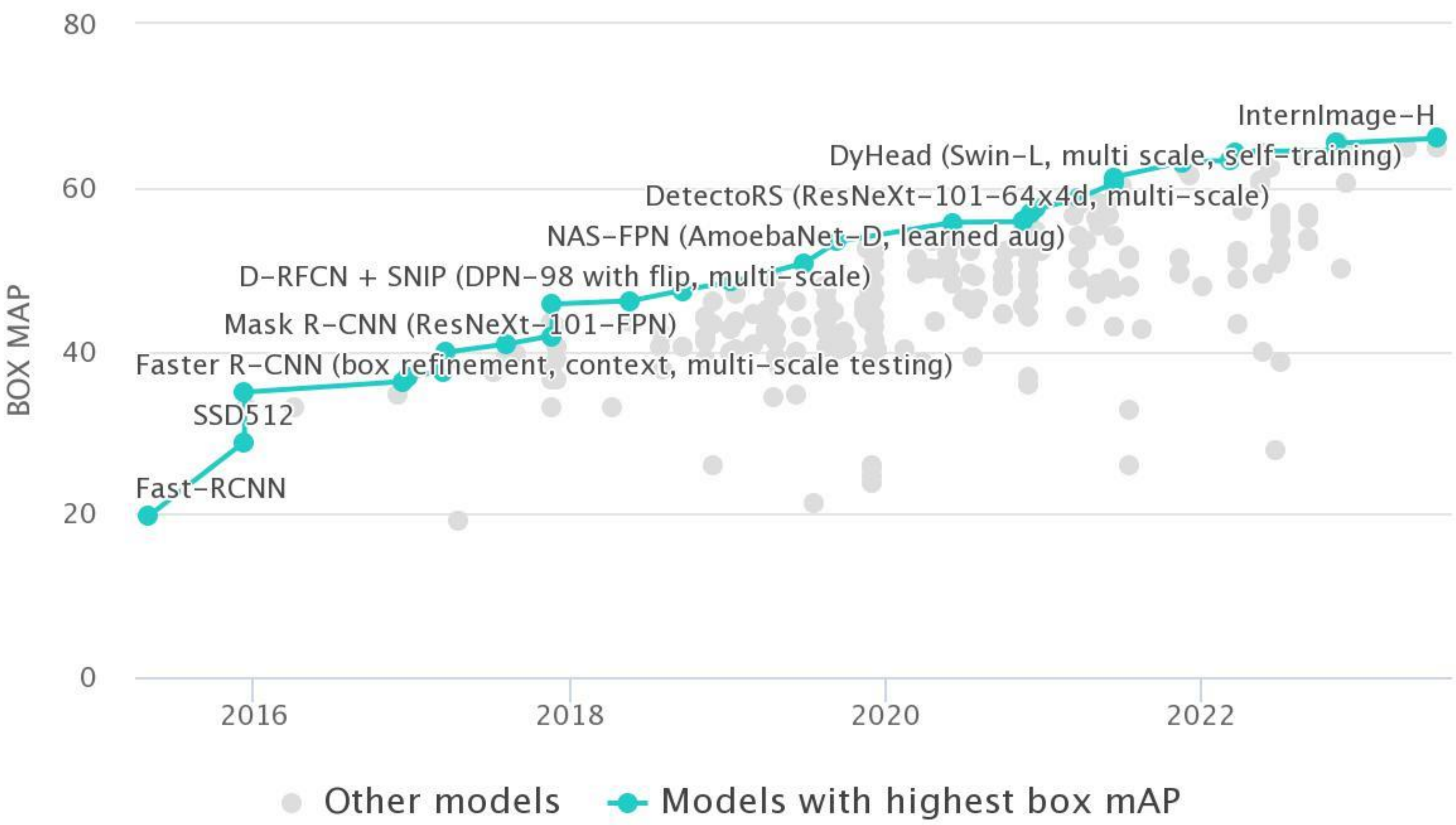
A decade of mad progress



Pascal-VOC-2012 semantic segmentation

https://paperswithcode.com/sota/semantic-segmentation-on-pascal-voc-2012

A decade of mad progress



MS-COCO object detection

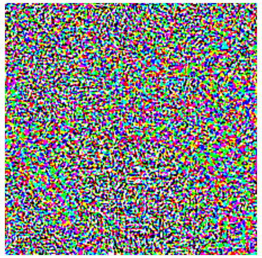
https://paperswithcode.com/sota/object-detection-on-coco

Yes, but

- Numerous errors even in controlled dev-test
- Many more under distribution shifts
- Extreme brittleness
- Possible absurd predictions

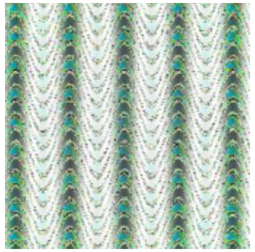
Panda

+ 0.07 x

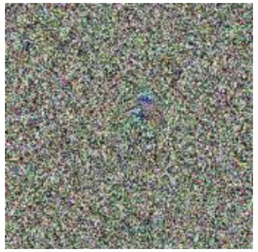


=

Gibbon



Peacock



Peacock

Starfish

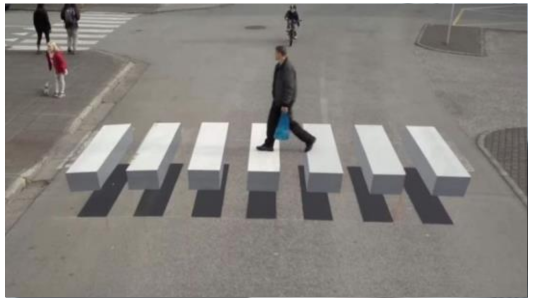
From intended to covered domain

Dataset defines the actual domain, often with limited coverage of:

- Rare pose/appearance of known objects, rare objects
- Rare, e.g. dangerous, scene configurations
- All sorts of perturbation, e.g., adverse conditions, sensor blocking





Expectations for real-world AI systems

Useful and safe models should

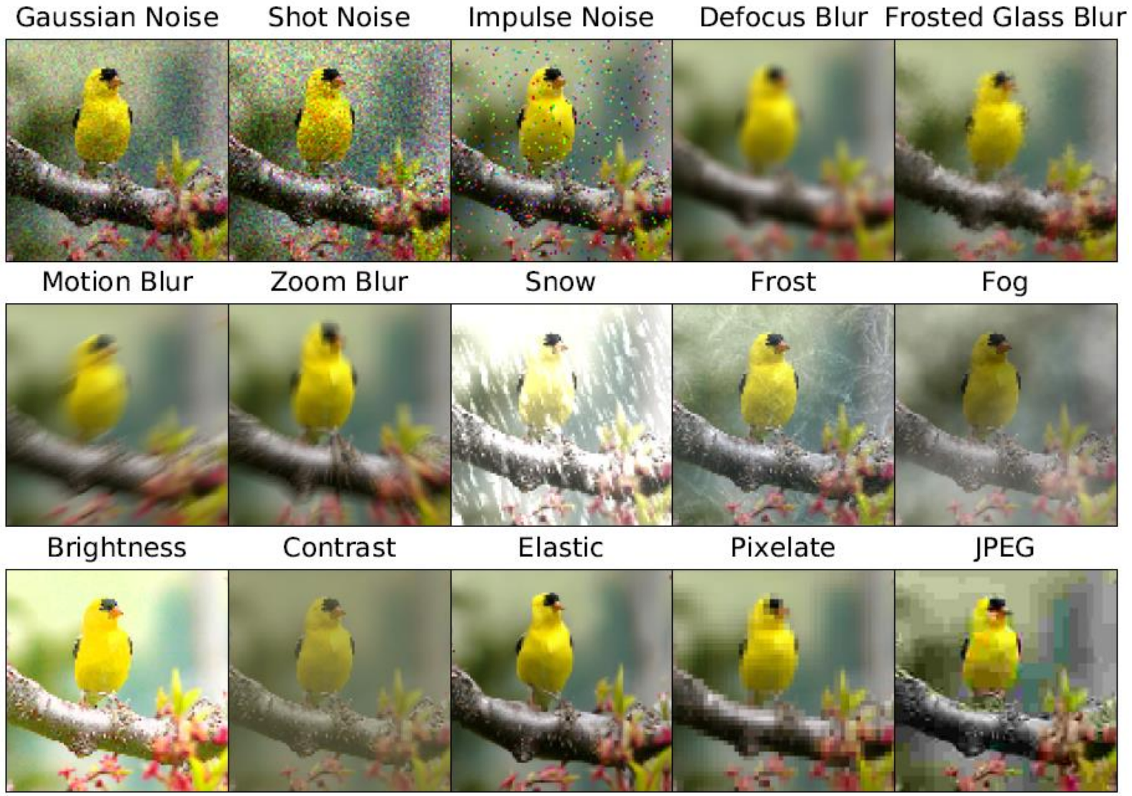
- Be **accurate** over *intended* domain, inc. corner cases
- Be **robust** to perturbations in-domain
- **Self-assess its confidence** for each prediction
- **Refuse to predict** if too uncertain, detect out-of-domain inputs
- **Adapt/generalize** to new domains or conditions

good dev-test accuracy does not suffice

All faces of runtime reliability should be assed and improved

Assessing robustness to corruption

- Various types/degrees of synthetic corruption on val/test data
- Measure their influence on model performance

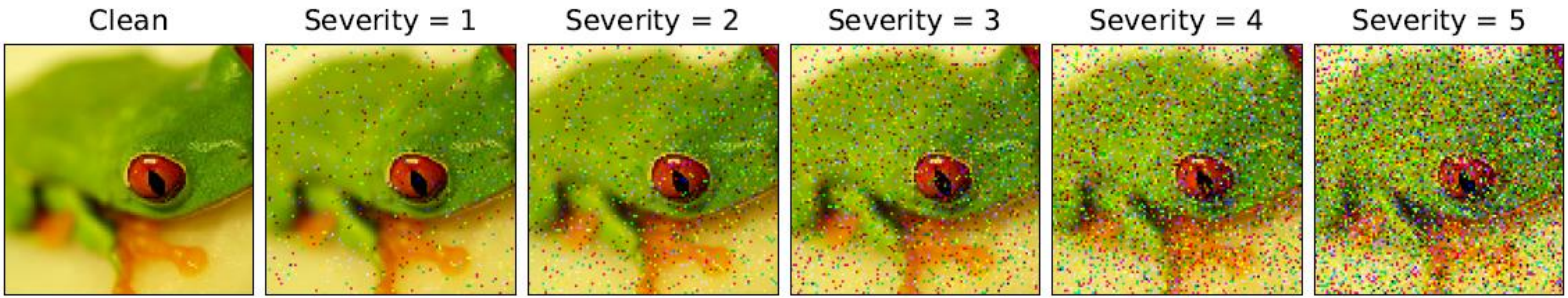


ImageNet-C

[Hendrycks ICLR 19]

Assessing robustness to corruption

- Various types/degrees of synthetic corruption on val/test data
- Measure their influence on model performance

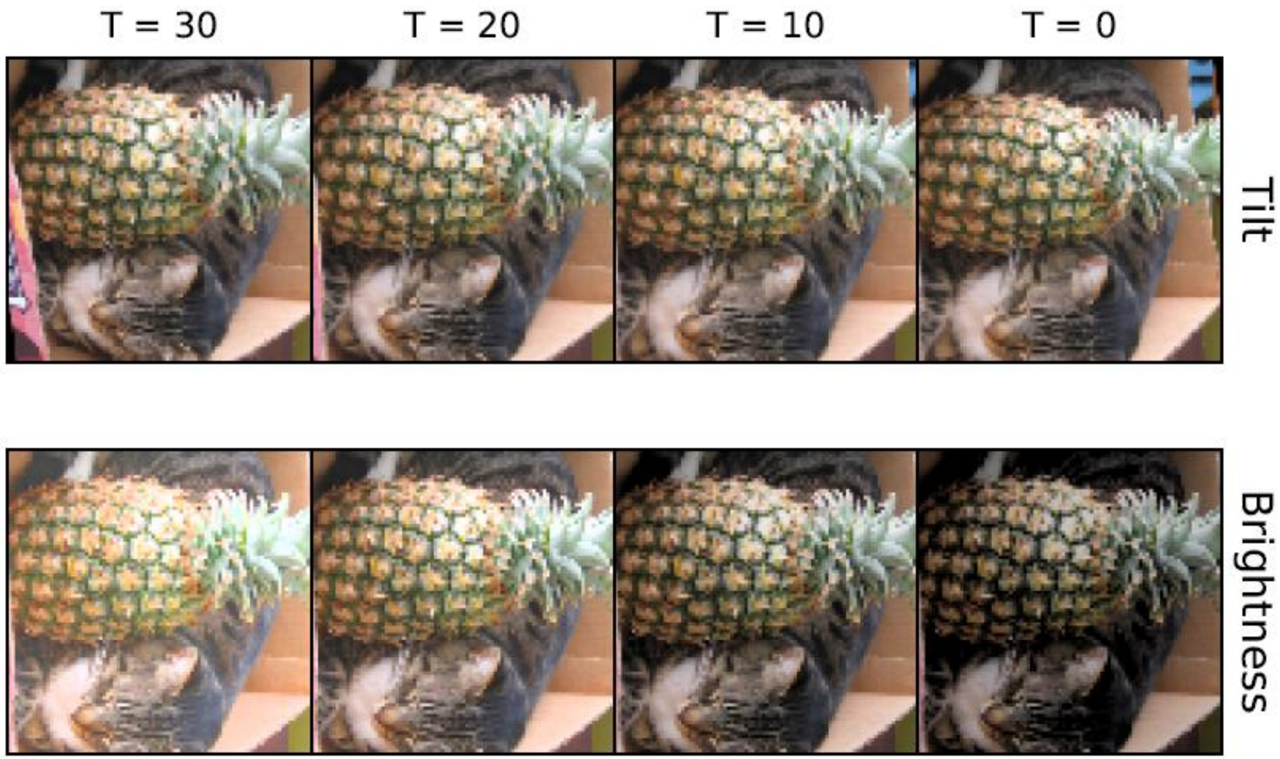


ImageNet-C

[Hendrycks ICLR 19]

Assessing robustness to perturbations

- Various types of gradual perturbations on val/test data
- Measure **invariance/covariance** of model w.r.t. them



ImageNet-P

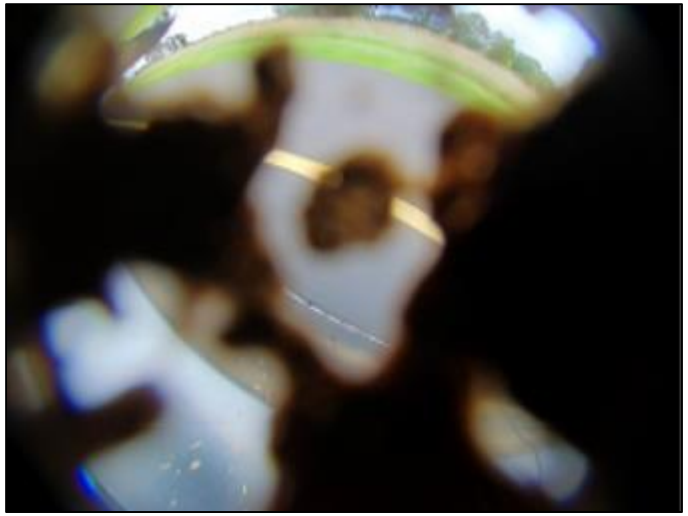
[Hendrycks ICLR 19]

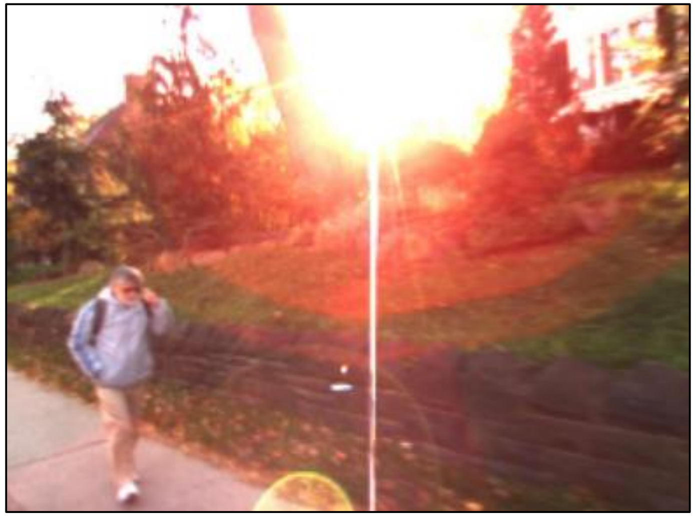
From synthetic to real perturbations

Synthetic perturbations: growing offer of robustness test datasets

- Objects: ImageNet-C,P,R
- Driving: Cityscapes foggy-, rainy-, -C, StreetHazards, fishyscapes, roadAnomaly21, roadObstacles21

Real perturbations: scattered across real datasets with little metadata





Why confidence prediction?

For development

- Help architecture design
- Guide annotation and training

For deployment

- Help validation
- Improve run-time reliability
- Help gain user’s trust

Why confidence prediction?

For development

- Help architecture design
- Guide annotation and training

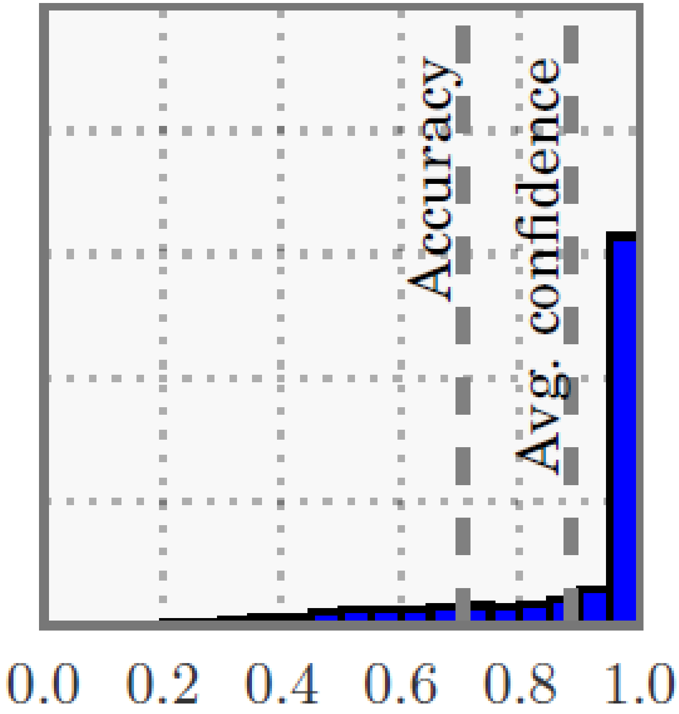
For deployment

- Help validation
- Improve run-time reliability
- Help gain user's trust

If very uncertain at run-time

- Inform downstream tasks
- Inform next time-step prediction
- Adapt sensor fusion, leverage redundancy
- Raise alarm
- Give control to another system…
- … or to human (if in the loop)
- Resort to emergency fallback

Max class score as confidence measure?

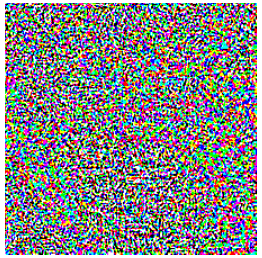


ResNet on Cifar 100

[Guo 2017]

[Goodfellow 2015]

+ 0.07 x

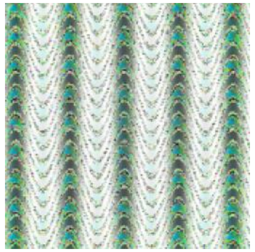


=

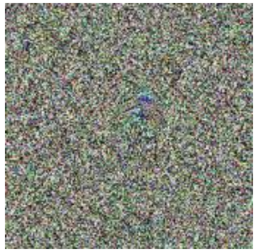
panda

gibbon

[Nguyen 2015]



peacock



peacock

starfish

max class score > 0.99

Roadmap to reliability

Assess model reliability

- Assess performance with a mix of criteria (not only Acc)
- Assess accuracy and robustness on realistic distribution shifts

evaluation

Improve model reliability

- Improve robustness to perturbations and OoD
- Adapt / generalize to new domains

models and training

Given input

- Predict confidence of model prediction
- Predict failure in classification
- Measure different types of uncertainties
- Detect if OoD (through model lens or not)
- Have a rejection option

run-time tools

Improve training data coverage inc. of corner cases

data dev cycle

ICCV 2023 Tutorial

🎭 The Many Faces of Reliability of Deep Learning for Real-World Deployment 🌍

Tuesday, October 3rd 2023, 08:30 - 13:00
Room S05
Paris, France

Schedule

08:30 - 08:50 **Setting the stage: from academic benchmarks to real-world situations** by **Patrick**

08:50 - 09:25 **Uncertainty estimation and next generation ensembles** by **Andrei**

09:25 - 10:20 **Calibration of Deep Neural Networks** by **Puneet**

10:20 - 10:40 **Break**

10:40 - 11:35 **Out-of-distribution detection** by **Sharon**

11:35 - 12:30 **Robustness and generalization under distribution shift** by **Dengxin** and **Tuan-Hung**

12:30 - 12:45 **Performance monitoring** by **Andrei**

12:45 - 13:00 **Closing remarks + Q&A** by **All**