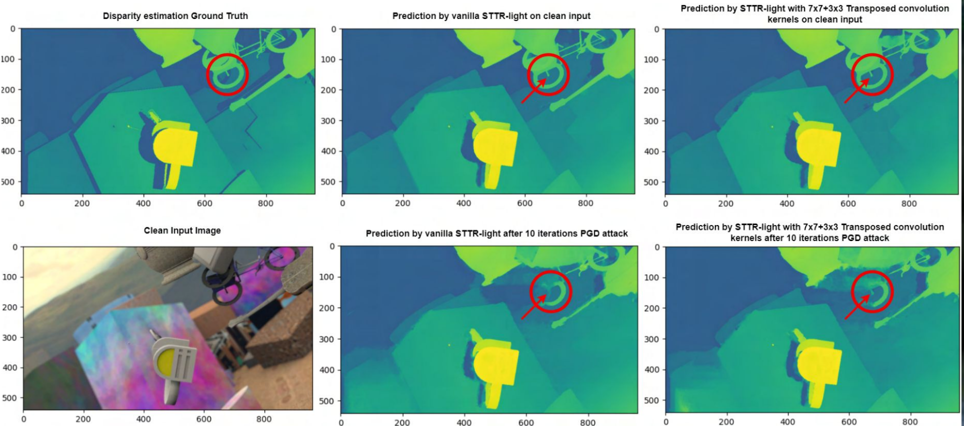
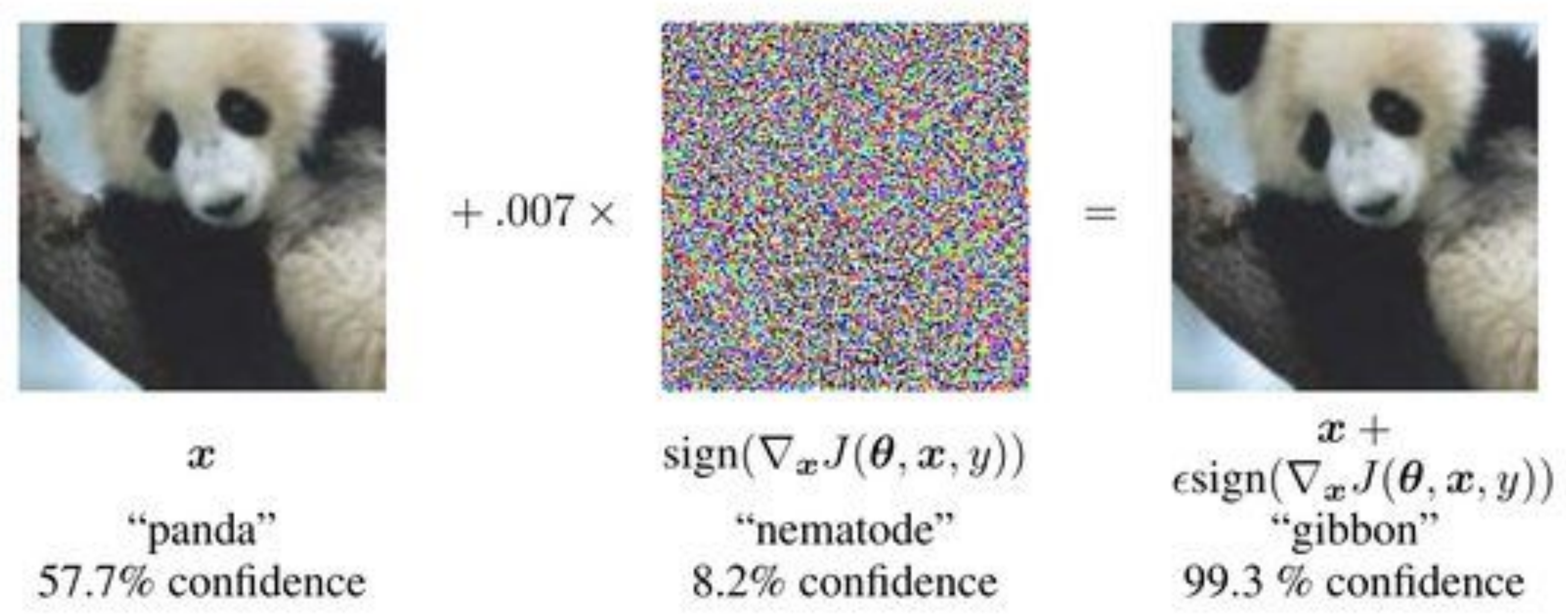
Team Projects:

Measuring Robustness for Pixel-wise prediction tasks

Machine Learning, DWS

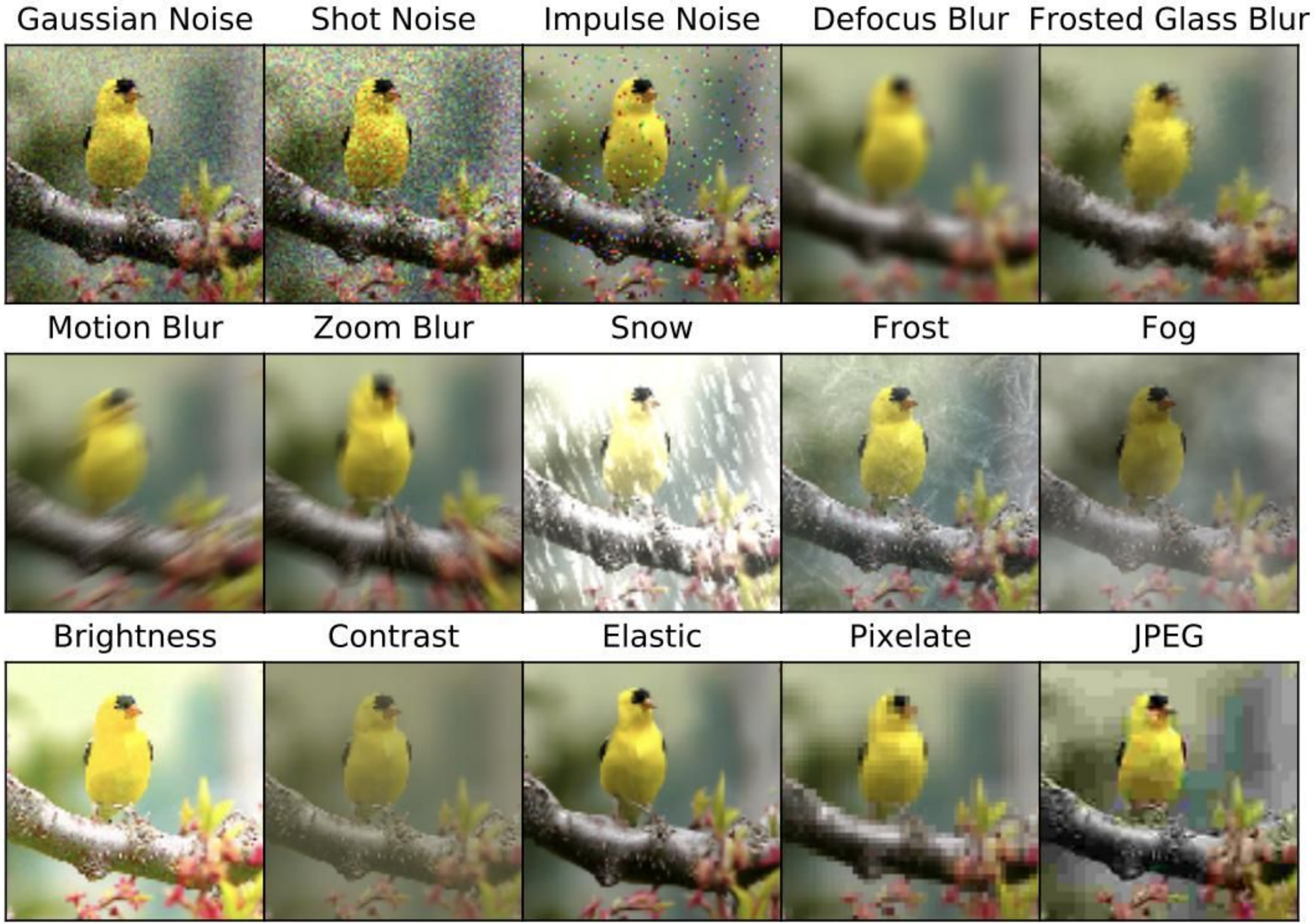


Adversarial Robustness



Source: Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Out-of-Distribution (OOD) Robustness



Source: Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Benchmarks exist for classification

arXiv
https://arxiv.org › cs

a standardized adversarial robustness benchmark

by F Croce · 2020 · Cited by 468 — Title:**RobustBench**: a standardized adversarial robustness benchmark ; Subjects: Machine Learning (cs.LG); Cryptography and Security (cs.CR); ...

ROBUSTBENCH Leaderboards Paper FAQ Contribute Model Zoo

ROBUSTBENCH

A standardized benchmark for adversarial robustness

The goal of **RobustBench** is to systematically track the *real* progress in adversarial robustness. There are already more than 3'000 papers on this topic, but it is still unclear which approaches really work and which only lead to overestimated robustness. We start from benchmarking common corruptions, ℓ_∞- and ℓ_2-robustness since these are the most studied settings in the literature. We use AutoAttack, an ensemble of white-box and black-box attacks, to standardize the evaluation (for details see our paper) of the ℓ_p robustness and CIFAR-10-C for the evaluation of robustness to common corruptions. Additionally, we open source the RobustBench library that contains models used for the leaderboard to facilitate their usage for downstream applications.

To prevent potential overadaptation of new defenses to AutoAttack, we also welcome external evaluations based on *adaptive attacks*, especially where AutoAttack flags a potential overestimation of robustness. For each model, we are interested in the best known robust accuracy and see AutoAttack and adaptive attacks as complementary.

News:

- **May 2022:** We have extended the common corruptions leaderboard on ImageNet with 3D Common Corruptions (ImageNet-3DCC). ImageNet-3DCC evaluation is interesting since (1) it includes more realistic corruptions and (2) it can be used to assess generalization of the existing models which may have overfitted to ImageNet-C. For a quickstart, click here. See the new leaderboard with ImageNet-C and ImageNet-3DCC here (also mCE metrics can be found here).
- **May 2022:** We fixed the preprocessing issue for ImageNet corruption evaluations: previously we used resize to 256x256 and central crop to 224x224 which wasn't necessary since the ImageNet-C images are already 224x224. Note that this changed the ranking between the top-1 and top-2 entries.

Up-to-date leaderboard based on 120+ models

Unified access to 80+ state-of-the-art robust models via Model Zoo

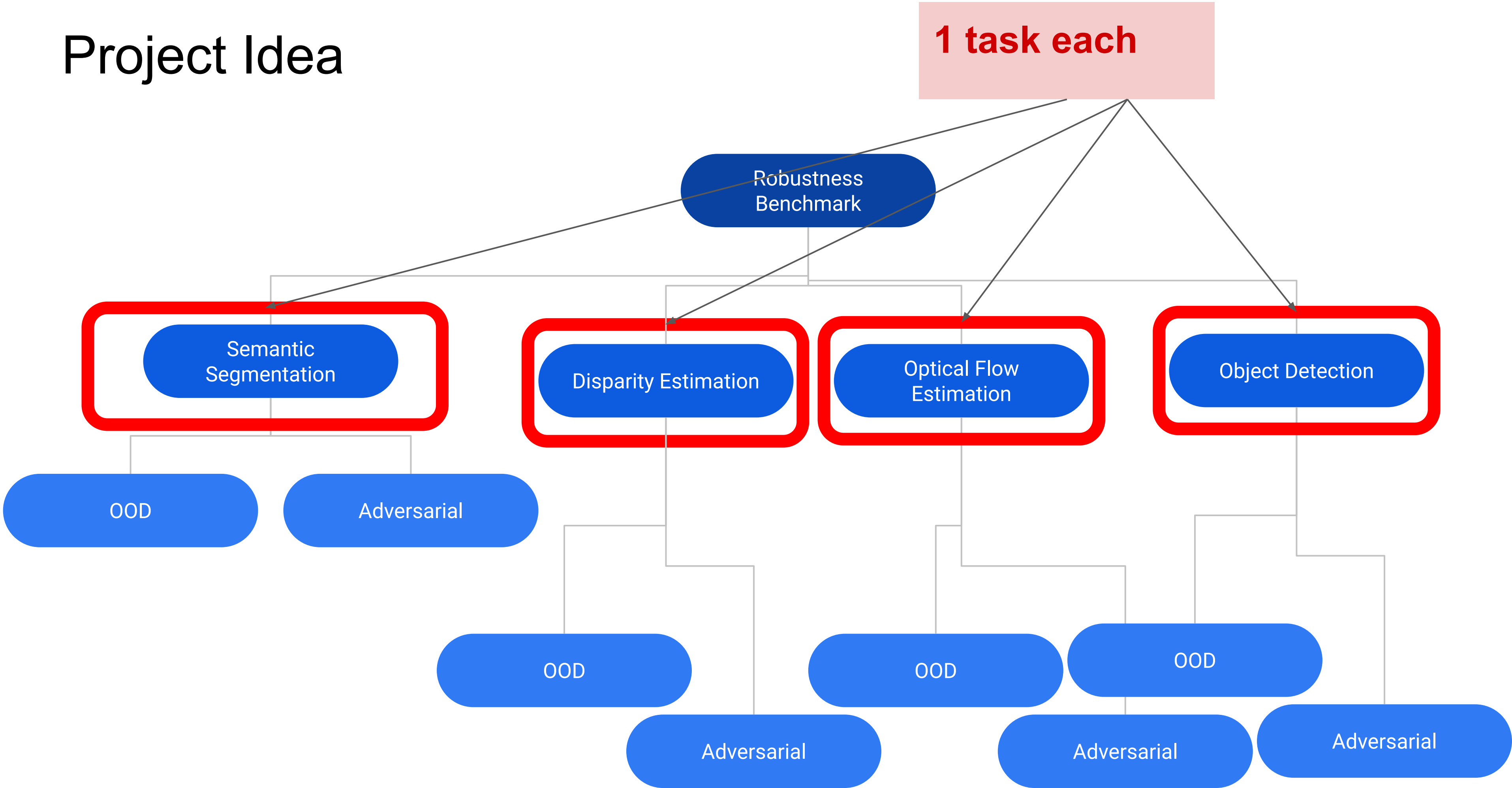
Model Zoo

Check out the available models and our Colab tutorials.

Analysis

Check out our paper with a detailed analysis.

Project Idea



Performance evaluation for OOD

- Common Corruptions: [Benchmarking Neural Network Robustness to Common Corruptions and Perturbations]
- 3D Common Corruptions:

  [3D Common Corruptions

  and Data Augmentation]

Semantic Segmentation

Architectures:

- UNet
- PSPNet
- DeepLabV3
- DeepLabV3+
- SegFormer
- InternImage
- Mask2Former
- DINO
- DINOv2
- ONE PEACE

Datasets:

- PASCAL VOC 2012
- Cityscapes
- ADE20K
- BDD

The current plan is to divide tasks for each vision task, by dataset.

Performance evaluation for Adversarial robustness

Test against following attacks:

- FGSM(iFGSM)
- PGD
- APGD
- CosPGD
- SegPGD
- AutoAttack

Epsilons:

- 1/255
- 2/255
- 4/255
- 8/255
- 12/255
- 25/255

Iterations:

- 1
- 3
- 5
- 10
- 20
- 40
- 100

Disparity Estimation

Architectures:

- UNet
- DispNet
- AutoDispNet
- STTR
- STTR-light
- CFNet
- HSMNet
- PSMNet
- GWCNet

Datasets:

- Flyingthing3D
- MPI Sintel
- KITTI 2015
- MIDDLEBURY_2014
- SCARED
- ETH 3D

The current plan is to divide tasks for each vision task, by set of architectures.

Optical Flow Estimation

Architectures:

- RAFT
- PWCNet
- GMANet
- SpyNet
- FlowNet2.0
- FlowFormer
- RPKNet

Additional Attacks:

- PCFA
- Adversarial Snow

Datasets:

- KITTI 2015
- MPI Sintel
- Spring

The current plan is to divide tasks for each vision task, by dataset.

Object Detection

Architectures:

- Co-DETR
- YOLOv5
- YOLOv7
- Faster R-CNN
- DETReg
- RetinaNet
- InternImage
- MogaNet
- EVA
- DINO
- DINOv2
- DETA

Datasets:

- PASCAL VOC 2007
- COCO
- ImageNet
- BDD
- KITTI 2012

The current plan is to divide tasks for each vision task, by dataset.

What can you expert to learn?

- Using pytorch
- State-of-the-art vision tasks and methods
- Robustness
- Building-up from a base code
- Potentially paper writing

What should you bring?

- Basic python skills
- Basic machine learning knowledge
- Learning attitude
- Collaborative nature

Expected Deadline?

Finish Experiments: May 05, 2024

Finish Paper Writing: June 01, 2024

Deadline might change according to NeuRIPS 2024 schedule.

If interested / Further Questions

Please contact: **Shashank Agnihotri**

Email id: **shashank.agnihotri@uni-mannheim.de**

Email subject: **Team Project 2024**

If interested:

Please include the vision tasks you are interested in the order of the priority

In case of many students, we are open to adding other vision tasks, such as depth estimation, panoptic segmentation etc.