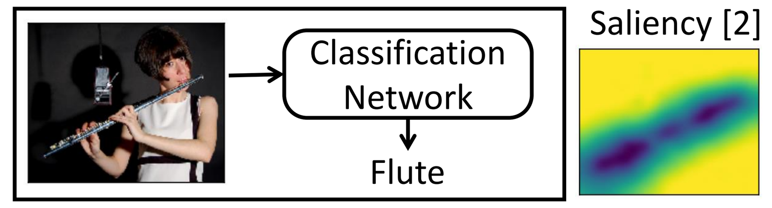
Grid Saliency for Context Explanations of Semantic Segmentation

Lukas Hoyer, Mauricio Munoz, Prateek Katiyar, Anna Khoreva, Volker Fischer
Bosch Center for Artificial Intelligence

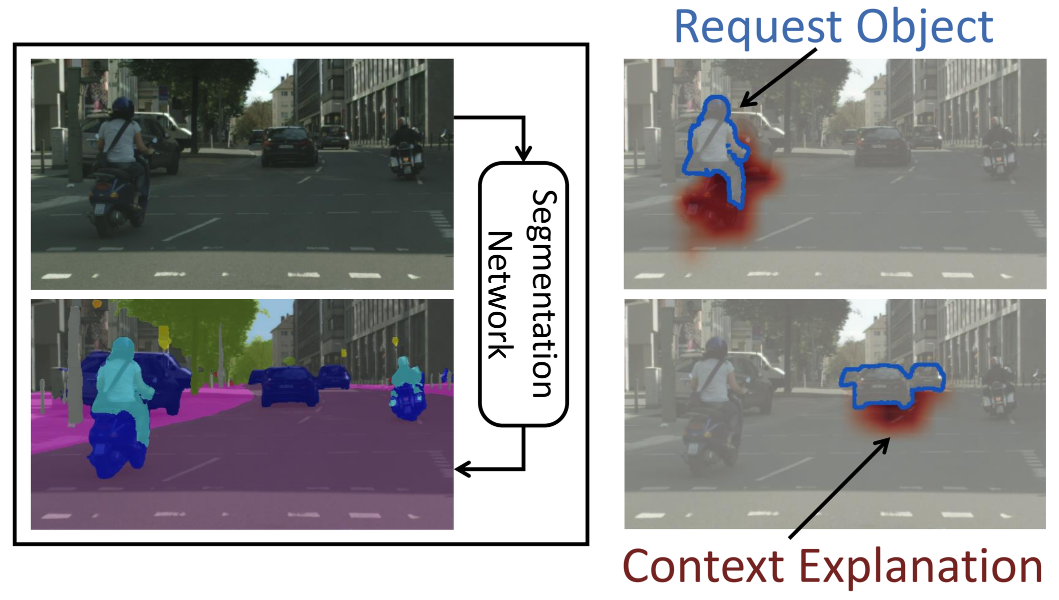
1 Motivation

Prior work: Saliency maps are used for visual explanations of neural network image classifications [1,2,3,4]



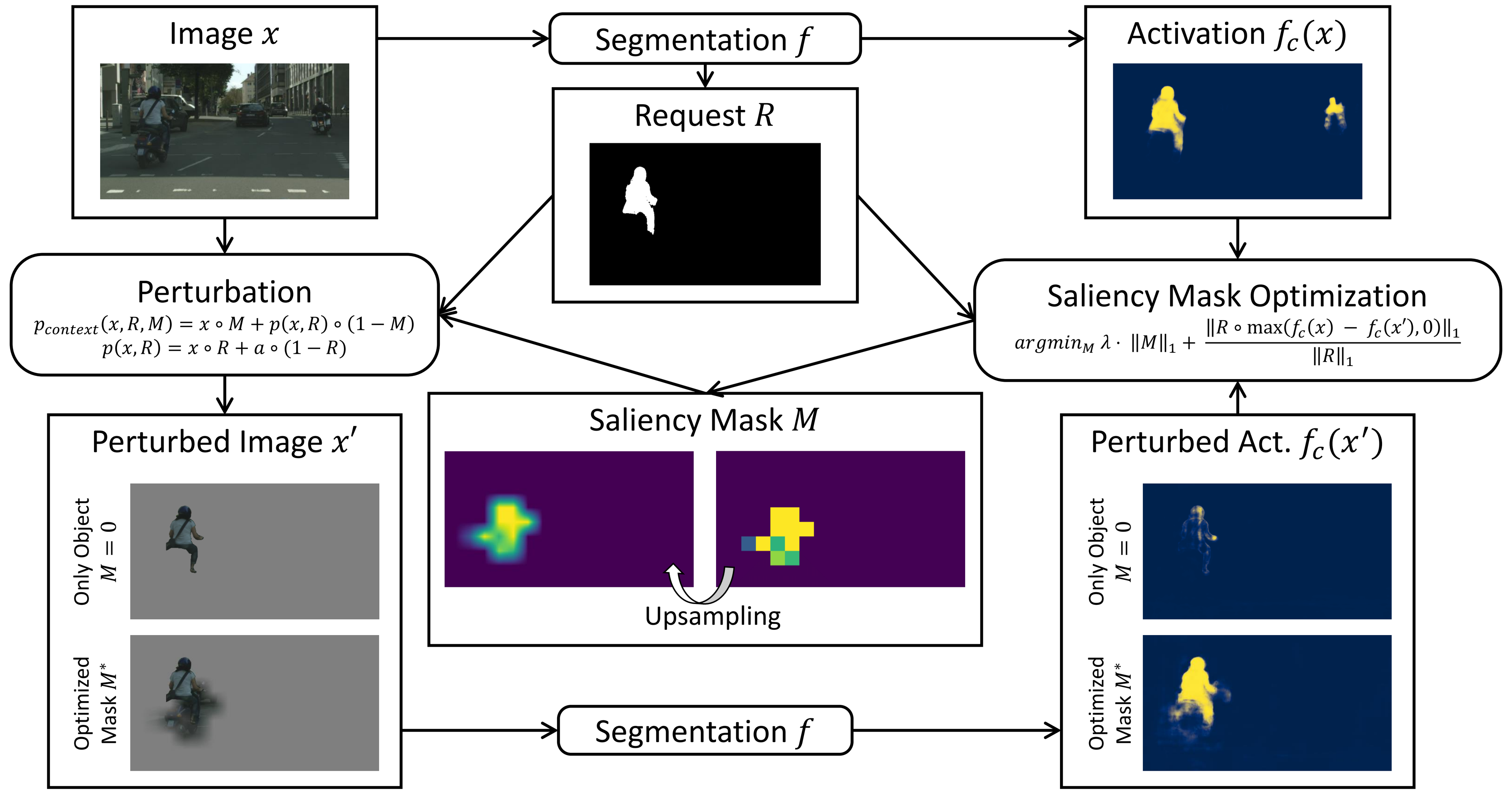
Our approach: Grid Saliencies for dense prediction tasks

Contributions:
- Context explanations for semantic segmentation
- Synthetic benchmark
- Real-world data analysis



2 Perturbation-Based Grid Saliency

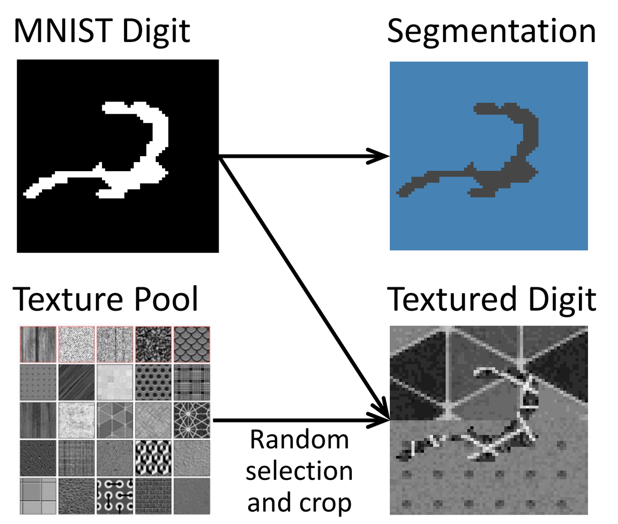
Context saliency map optimization
- Perturb as much context as possible
- Retain network predictions inside a request region



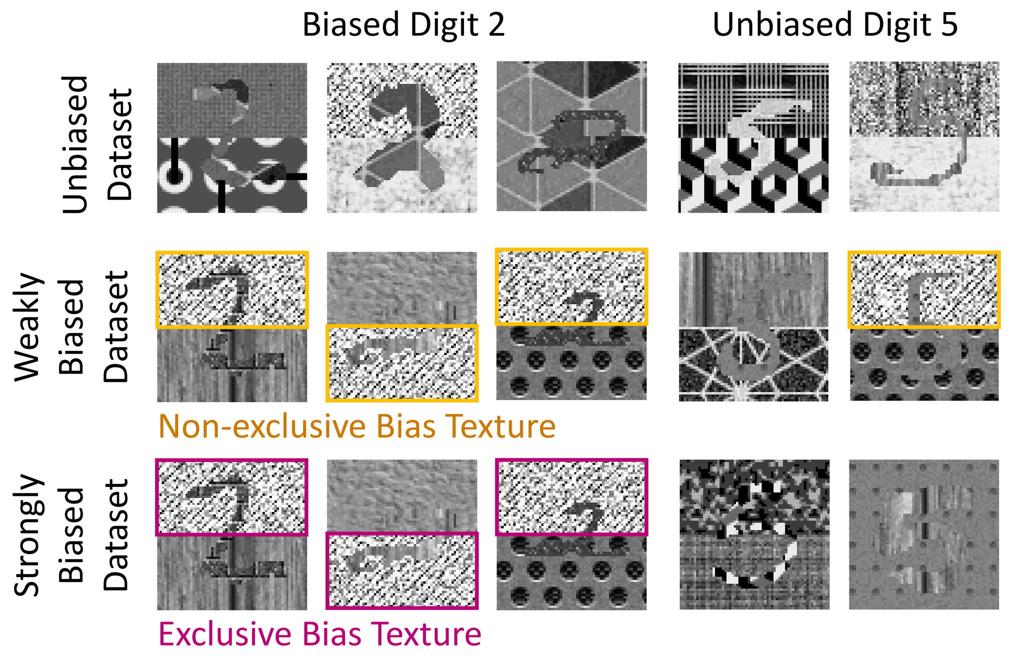
3 Synthetic Data for Benchmarking Saliencies

Generation of dataset with induced bias

MNIST digits are combined with random fore- and background textures

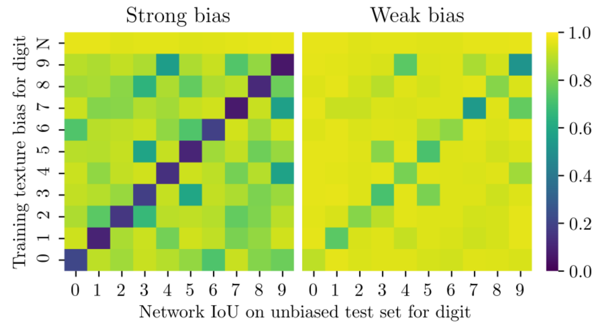


Context bias is induced into dataset by choosing the same background texture for a specific digit

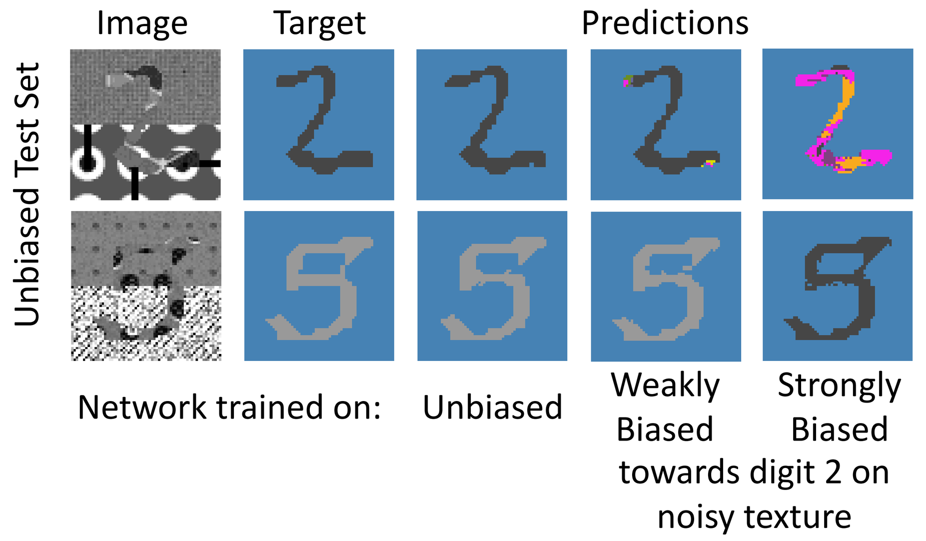


Has the network picked up the bias?

Test a potentially biased network on an unbiased dataset

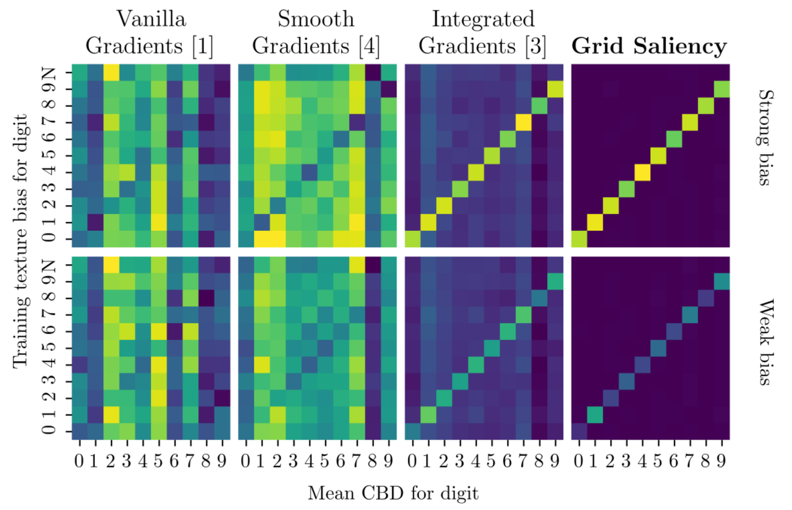


Drop in segmentation IoU for biased digit
→ Network has picked up the bias



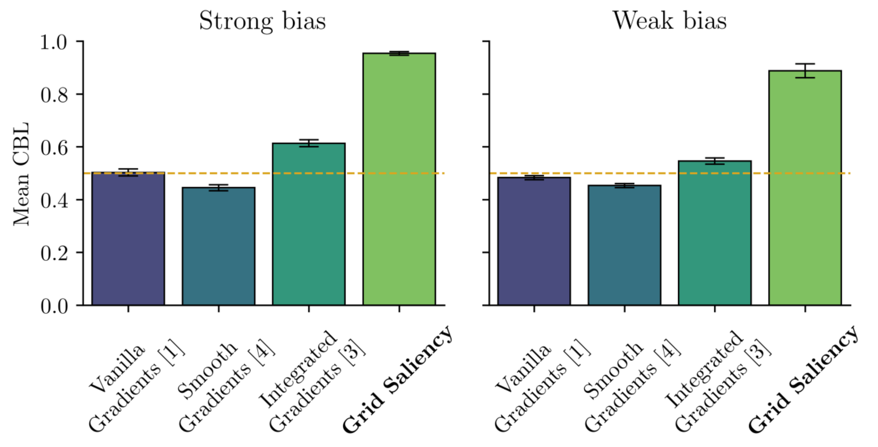
Can the saliency detect biased samples and localize the bias?

Measure $CBD = \frac{\textit{salient context}}{\textit{context}}$ averaged over 5 bias textures and 5 datasets



High grid saliency on biased digits
→ Biased samples can be detected

Measure $CBL = \frac{\textit{biased salient context}}{\textit{salient context}}$ averaged over 10 bias digits, 5 bias textures, and 5 datasets

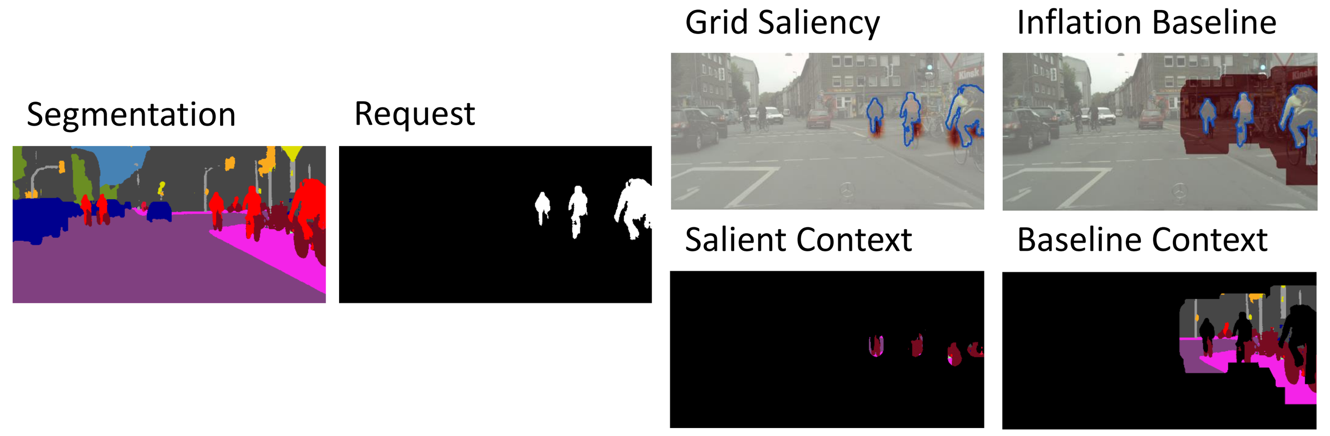


High grid saliency on biased context
→ Bias can be localized

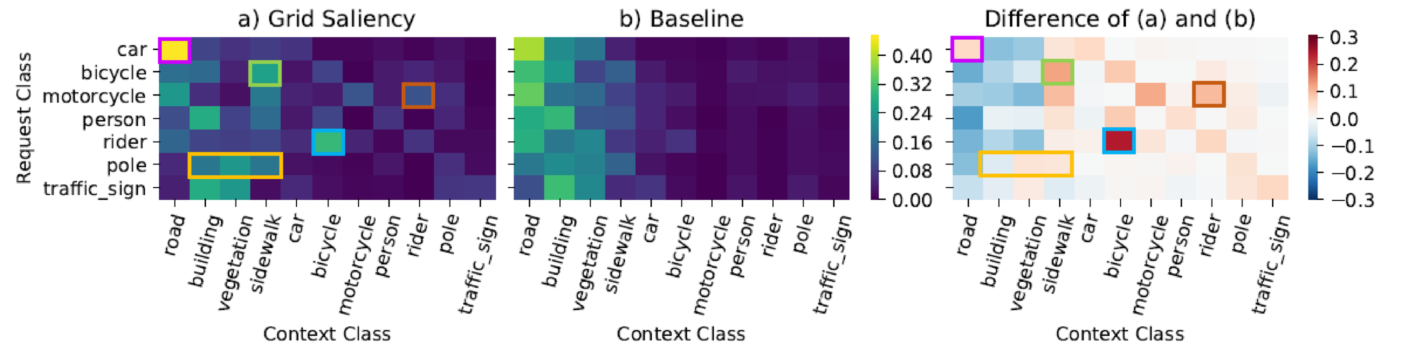
4 Grid Saliency on Cityscapes

Statistics on context explanations

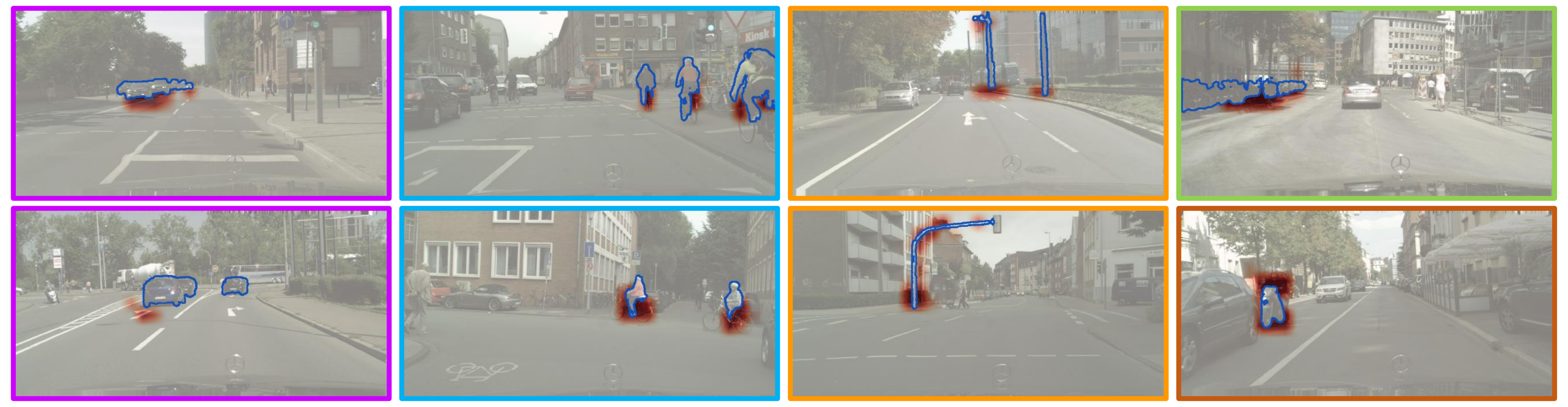
- Analysis of salient context over the Cityscapes validation set
- Relative amount of salient context pixels for different request classes
- Baseline: inflated object instances



Exemplary Findings:
- Road for car
- Bicycle for rider
- Vegetation for pole
- Sidewalk for bicycle
- Rider for motorcycle



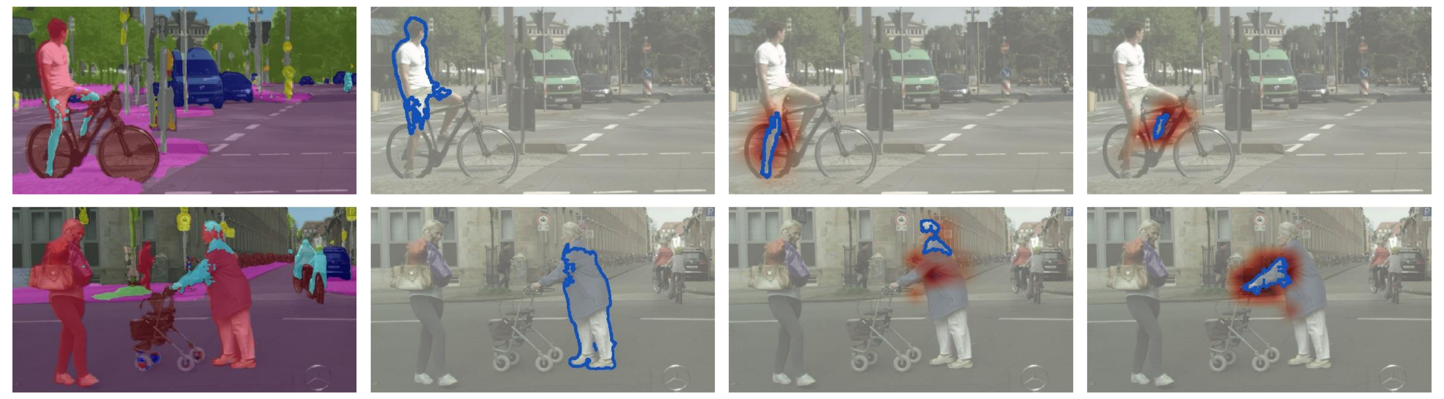
Examples of context explanations



Error case analysis

Rider classified as pedestrian → bicycle not salient

Pedestrian classified as rider → arm pose salient



5 References

[1] Karen Simonyan et al. Deep inside convolutional networks: Visualising image classification models and saliency maps. In International Conference on Learning Representations, 2013.

[2] Ruth Fong et al. Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the IEEE International Conference on Computer Vision, 2017.

[3] Mukund Sundararajan et al. Axiomatic attribution for deep networks. In International Conference on Machine Learning, 2017.

[4] Daniel Smilkov et al. Smoothgrad: removing noise by adding noise. arXiv:1706.03825, 2017.