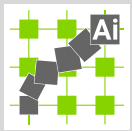
Safety and Security in Deep Neural Networks

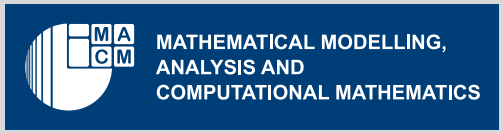
Matthias Rottmann and Hanno Gottschalk

joint work with R. Chan, K. Maag, P. Colling and M. Schubert

Arbeitsgruppe Angewandte Informatik,
Bergische Universität Wuppertal

Dortmund Data Science Center Kolloquium





Current Projects

Chasing the Errors of AI

Interdisciplinary Group with Hanno Gottschalk (stochastics)

Projects:

Uncertainty Quantification and Performance / Prediction Quality Estimates for

- Image Classification
- Semantic Segmentation (R. Chan, K. Maag, Volkswagen)
- Object Detection (M. Schubert, FIS.NRW)

Applications of this

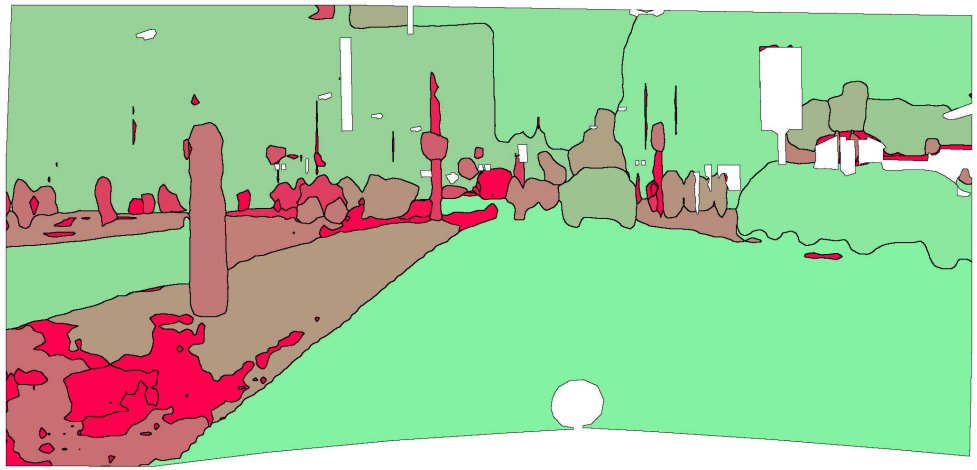
- Active Learning for Semantic Segmentation (P. Colling, Aptiv)
- Active Learning for Object Detection (M. Schubert, FIS.NRW)

Semantic Segmentation of Street Scenes

- can be learned by convolutional neural networks
- decision making is extremely intransparent $\rightarrow$ we work on methods for rating predictions (performance estimates)

Performance Estimates / Rating Predicted Segments

IoU

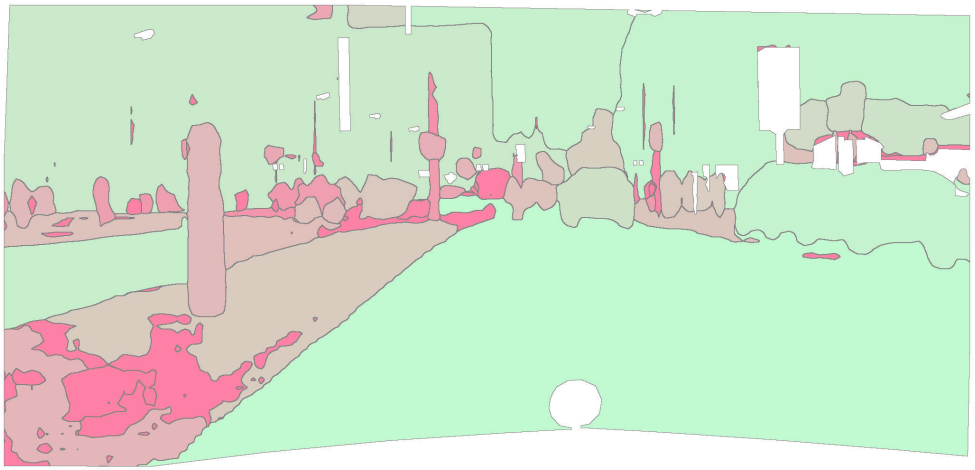


ground truth

prediction

Performance Estimates / Rating Predicted Segments

IoU



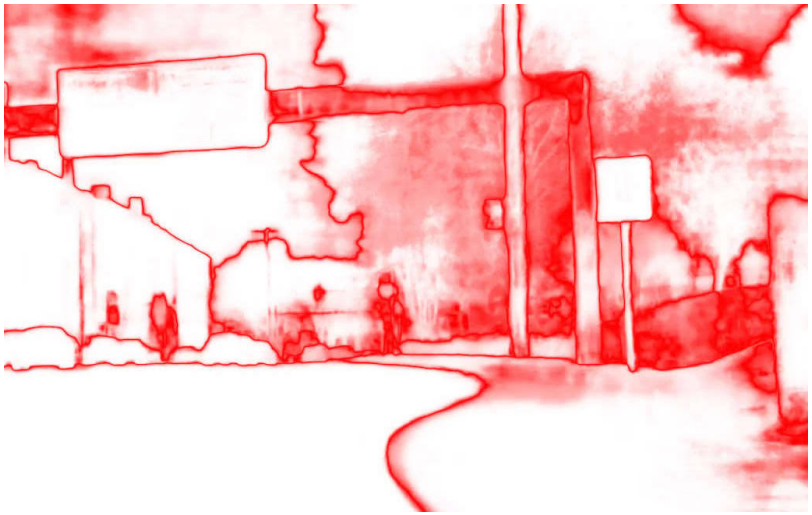
Goal: reliable prediction of the IoU

ground truth is not available at runtime

prediction

Segment-wise Aggregated Metrics and Performance Measures

Observations for low quality predictions:



segment filling dispersion

fuzzy/fractal shapes

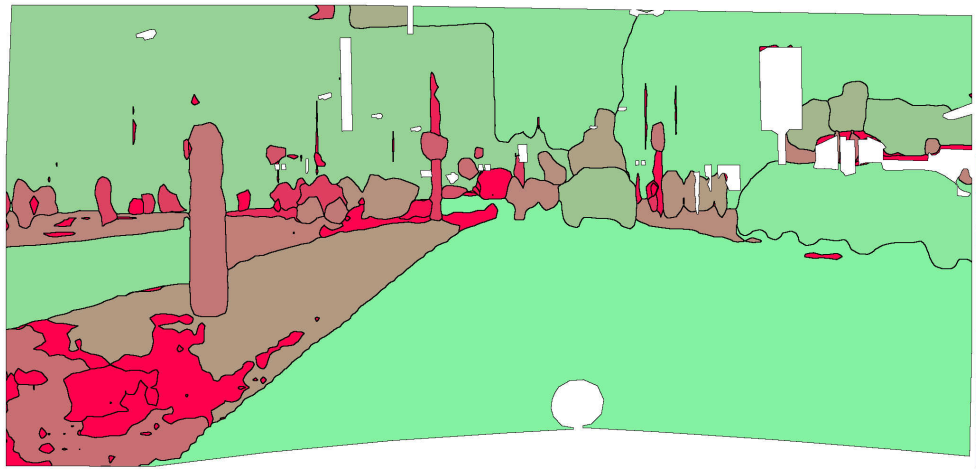
$\rightarrow$ construct metrics for quantification

$\rightarrow$ use them to predict the IoU

MetaSeg: Linear Regression with IoU

IoU

predicted IoU

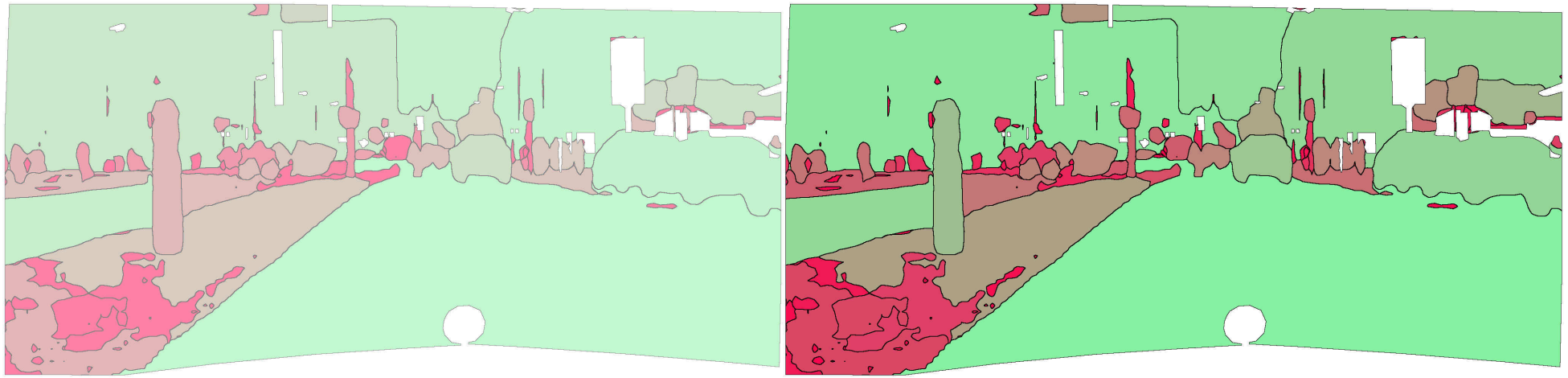


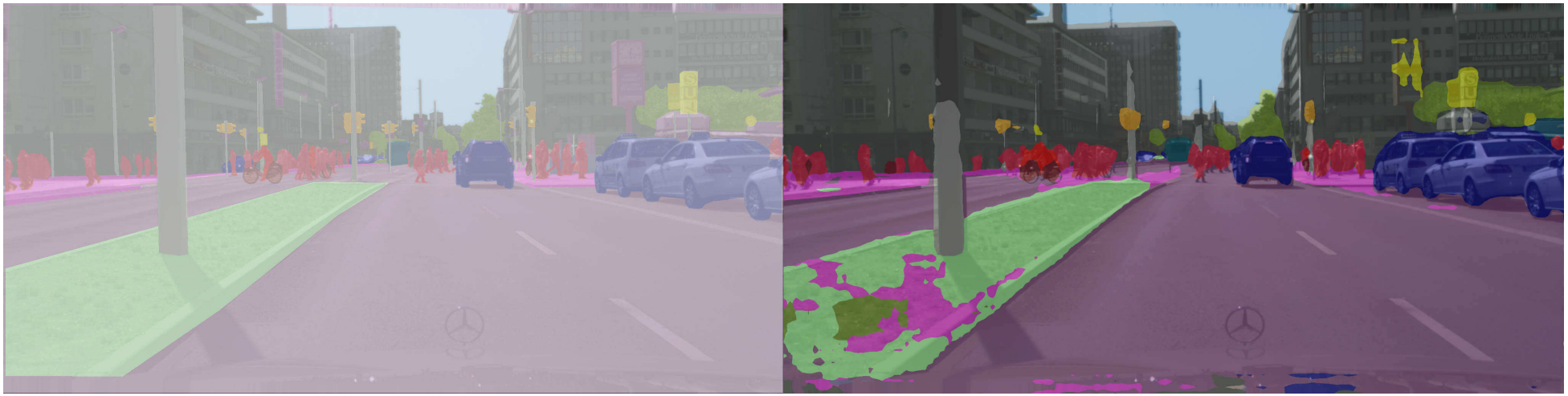
ground truth

prediction

MetaSeg: Linear Regression with IoU

predicted IoU





prediction

Current Projects

Further Projects:

- Decision Rules in Semantic Segmentation and their Ethical Consequences (R. Chan, Volkswagen)
- Detection of Adversarial Attacks on Image Classifiers (M. Peyron)

Upcoming:

- Algorithm Development for Meta Learning (AutoML)
- Prediction Quality Estimates as Additional Loss Functions
- Uncertainty Quantification for DL with Lidar/Radar Data
- Generative Adversarial Networks in Computer Simulations
- ...