Intelligent Systems Laboratory

uc3m

Stereo-Vision Based Convolutional Networks for Object Detection in Driving Environments

Carlos Guindel (cguindel@ing.uc3m.es) , David Martín, José M. Armingol

Sixteenth International Conference on Computer Aided Systems Theory
Intelligent Transportation Systems and Smart Mobility Workshop

Intelligent Systems Laboratory · Universidad Carlos III de Madrid

Outline

1. Introduction
2. Object detection pipeline
3. Proposed approach
4. Results
5. Conclusions

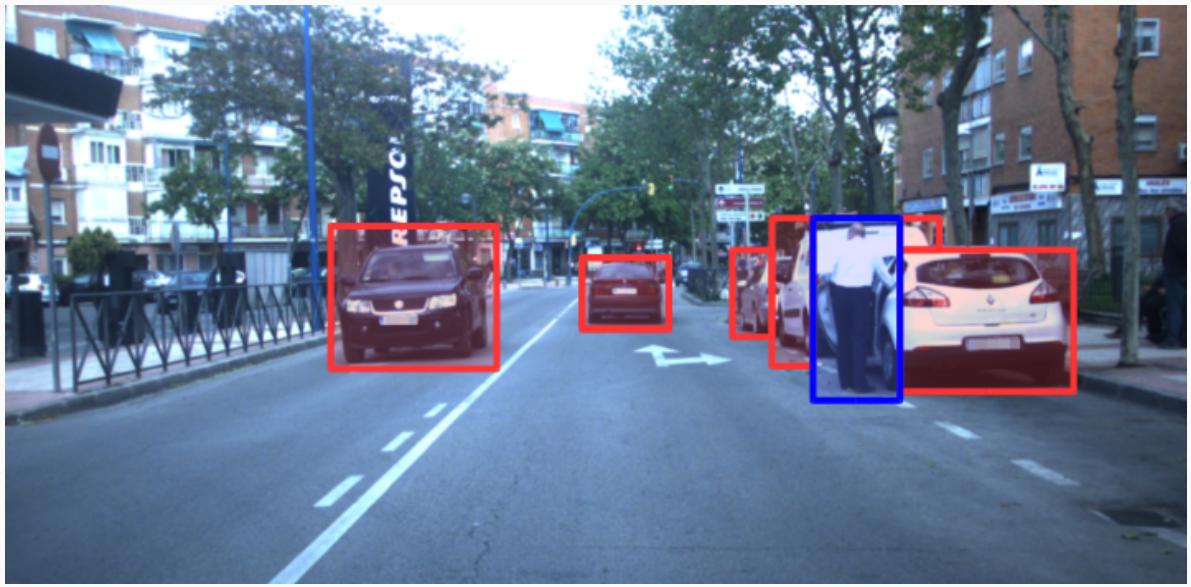
Introduction

Object detection in driving environments

- Modern Advanced Driver Assistance Systems (ADAS) and autonomous vehicles rely on robust **object detection**.

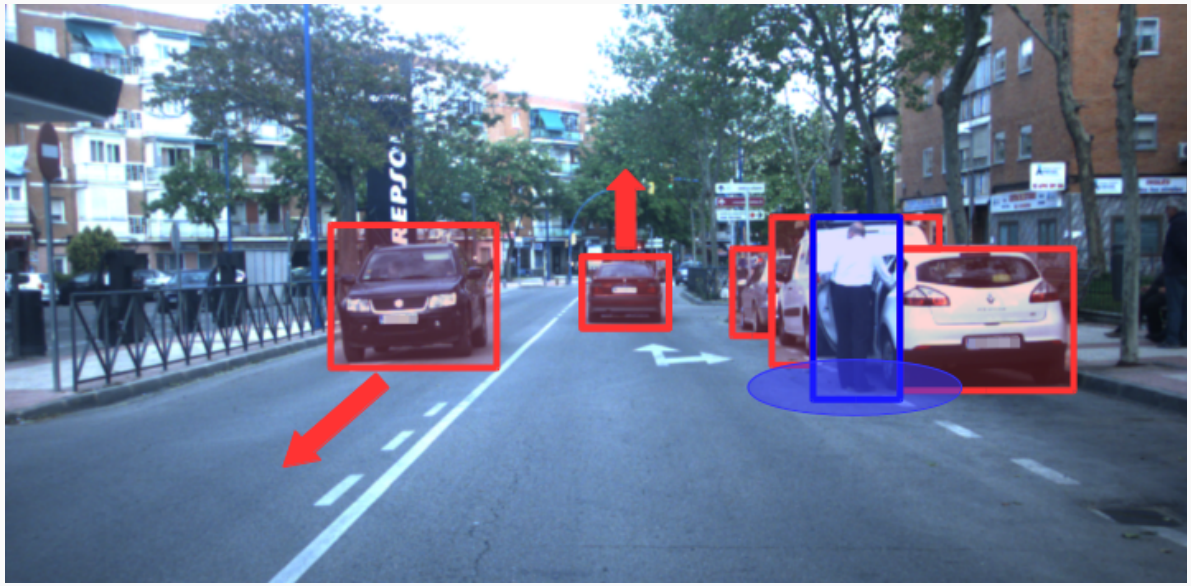
Object detection in driving environments

- Modern Advanced Driver Assistance Systems (ADAS) and autonomous vehicles rely on robust **object detection**.
- A full understanding of the traffic scene requires to correctly **classifying** the obstacles.



Object detection in driving environments

- Modern Advanced Driver Assistance Systems (ADAS) and autonomous vehicles rely on robust **object detection**.
- A full understanding of the traffic scene requires to correctly **classifying** the obstacles.
- Different agents are expected to exhibit different **behaviors**.



Object detection in driving environments

Driving environments

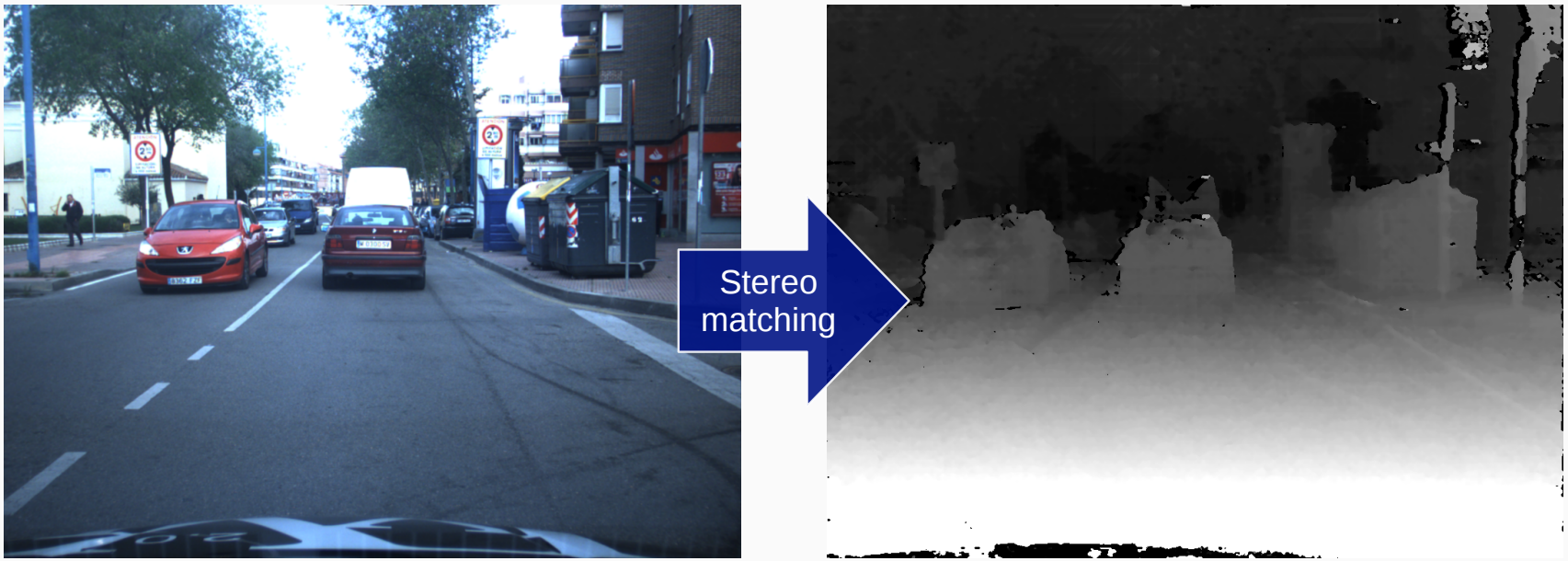
- Appearance changes
- Pose diversity
- Severe occlusions
- Background clutters

Stereo-vision based methods

- Stereo-vision systems are frequently used in automotive applications.

Stereo-vision based methods

- Stereo-vision systems are frequently used in automotive applications.
- They provide the same appearance information as monocular approaches, but also spatial reasoning.



Stereo-vision based methods

- Stereo-vision systems are frequently used in automotive applications.
- They provide the same appearance information as monocular approaches, but also spatial reasoning.

Common stereo-based approaches for obstacle detection

- Occupancy maps
- Digital elevation maps
- Scene flow
- Geometry-based clustering

State-of-art object detection methods

- **Convolutional Neural Networks** (deep learning) have recently emerged as a robust object detection paradigm.

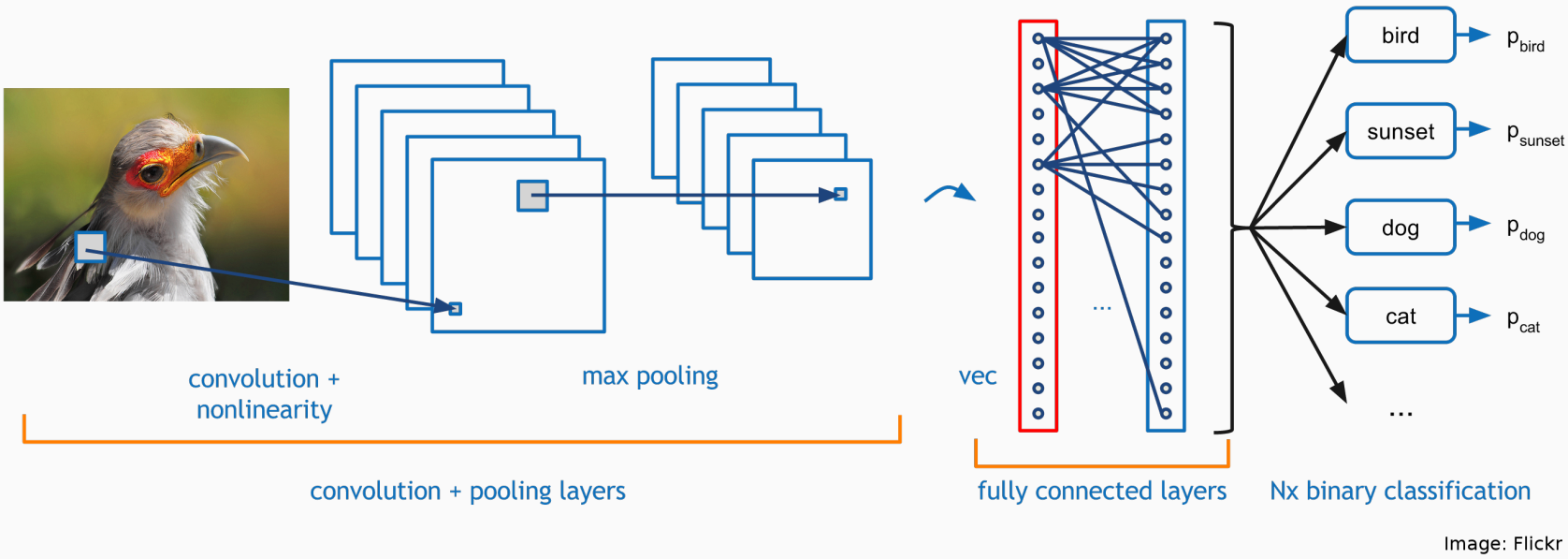


Image: Flickr

State-of-art object detection methods

- **Convolutional Neural Networks** (deep learning) have recently emerged as a robust object detection paradigm.
- **Traffic environments**: KITTI Object Detection Benchmark results are dominated by methods based on Convolutional Neural Networks (CNNs).

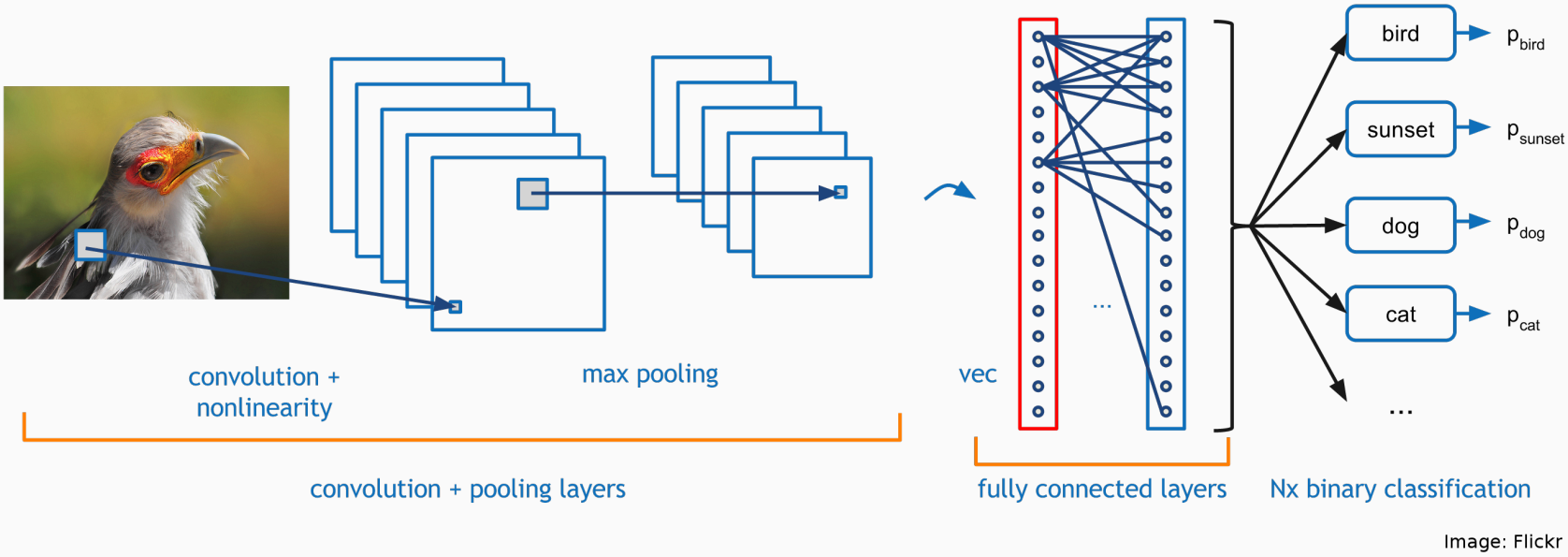
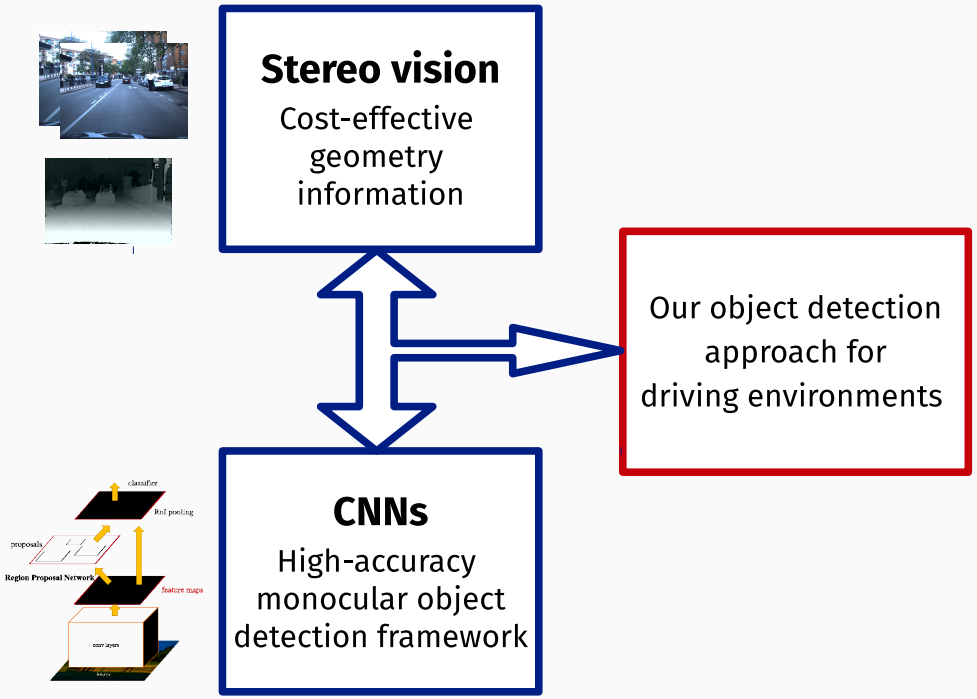


Image: Flickr

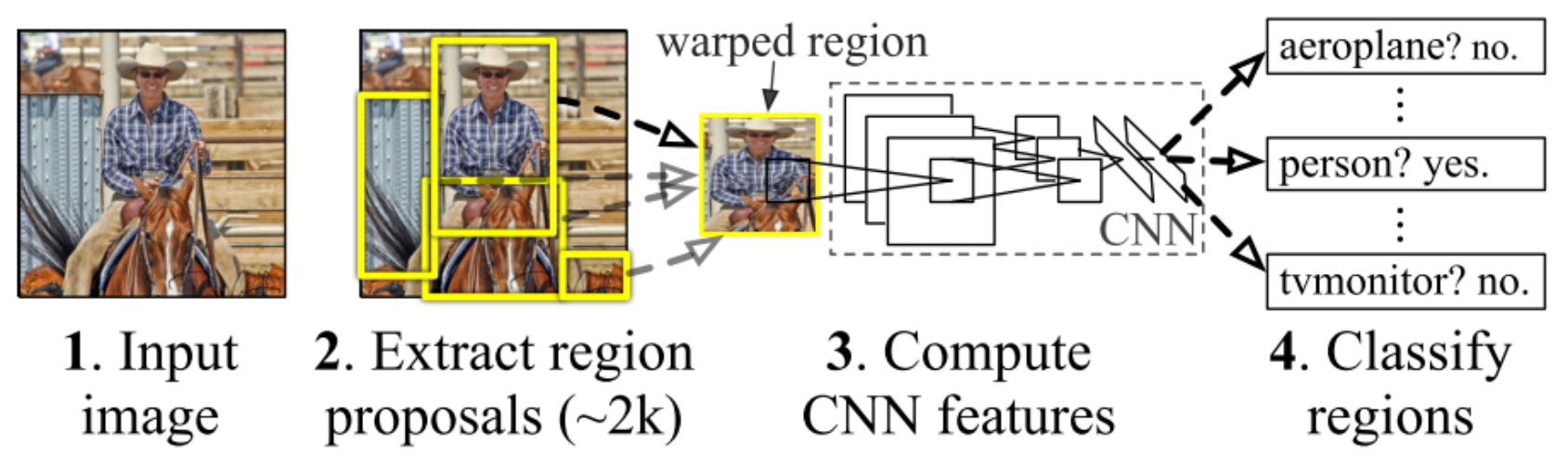
Our idea

- To integrate stereo information into a CNN object detection framework to make it more suitable for driving environments.



Object detection pipeline

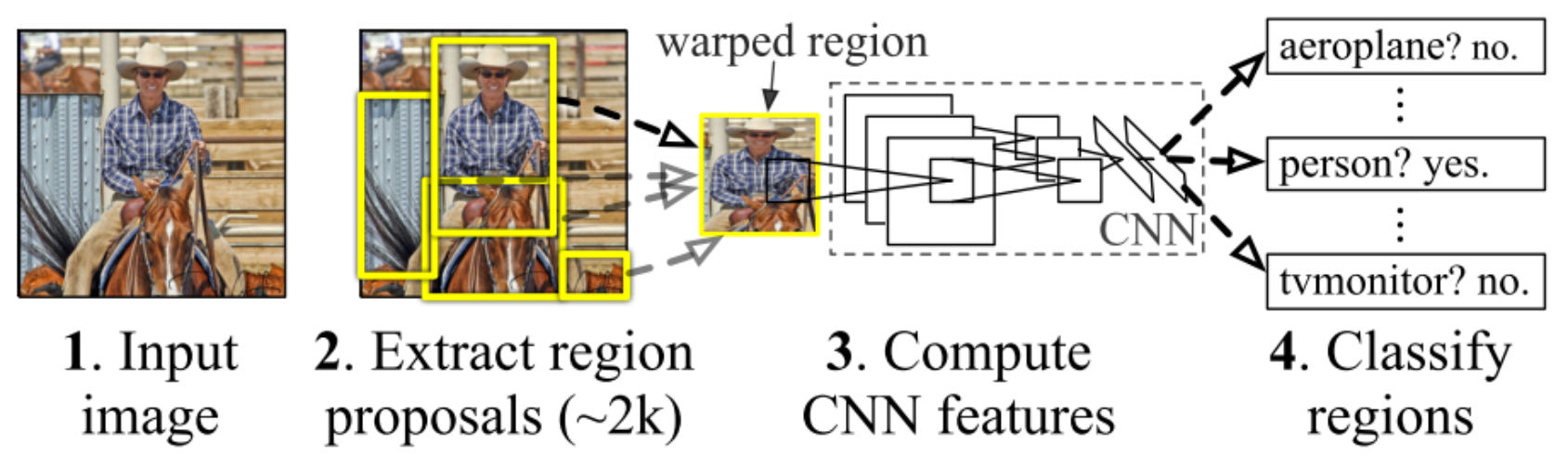
R-CNN framework



- R-CNN is a state-of-art CNN-based object detection pipeline.

R. Girshick, J. Donahue, et al., "Region-based Convolutional Networks for Accurate Object Detection and Segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 1, pp. 142–158, 2016.

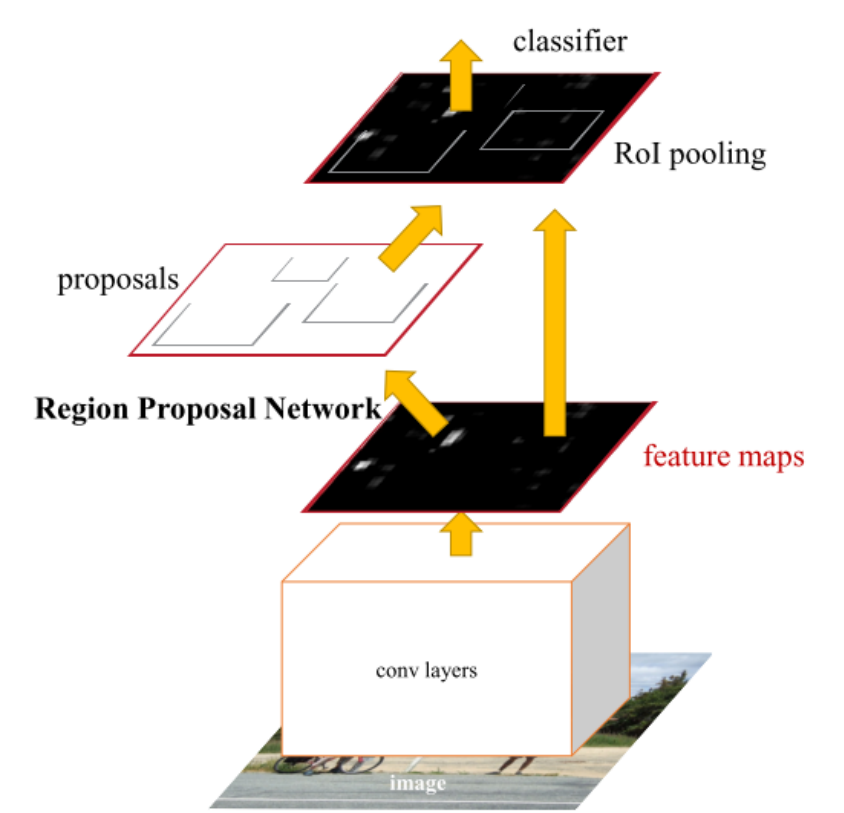
R-CNN framework



- R-CNN is a state-of-art CNN-based object detection pipeline.
- Extracts CNN features from previously selected regions of interest.

R. Girshick, J. Donahue, et al., "Region-based Convolutional Networks for Accurate Object Detection and Segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 1, pp. 142–158, 2016.

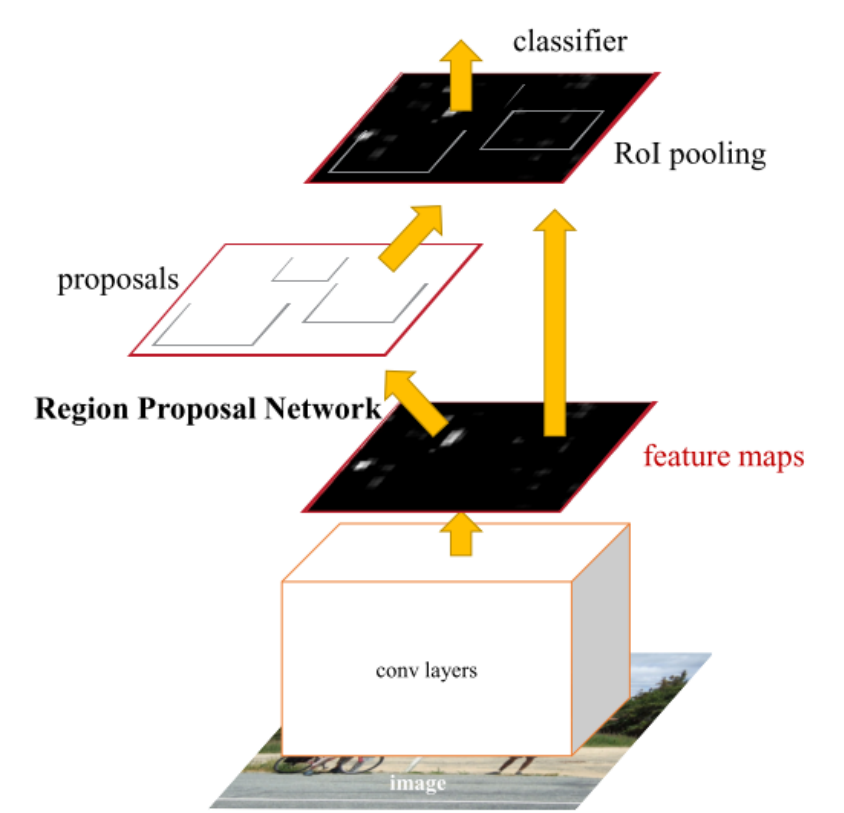
Faster R-CNN framework



- Faster R-CNN aims to be an end-to-end detection pipeline.

S. Ren, K. He, et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Adv. in Neural Information Processing Systems (NIPS), 2015.

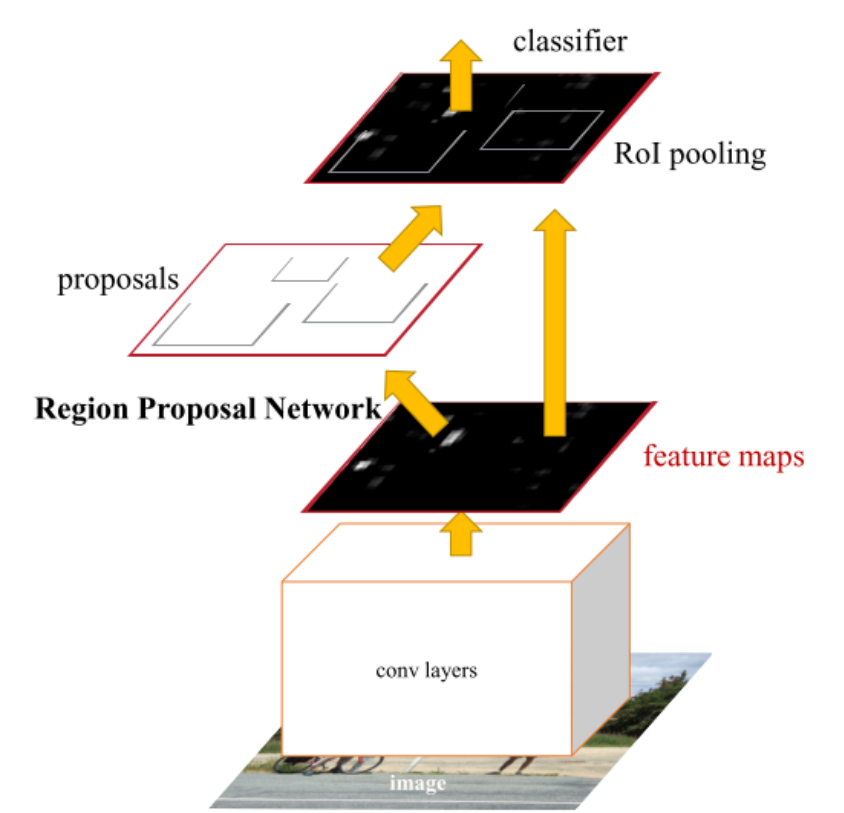
Faster R-CNN framework



- Faster R-CNN aims to be an end-to-end detection pipeline.
- Regions are proposed by a **Region Proposal Network (RPN)**

S. Ren, K. He, et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Adv. in Neural Information Processing Systems (NIPS), 2015.

Faster R-CNN framework

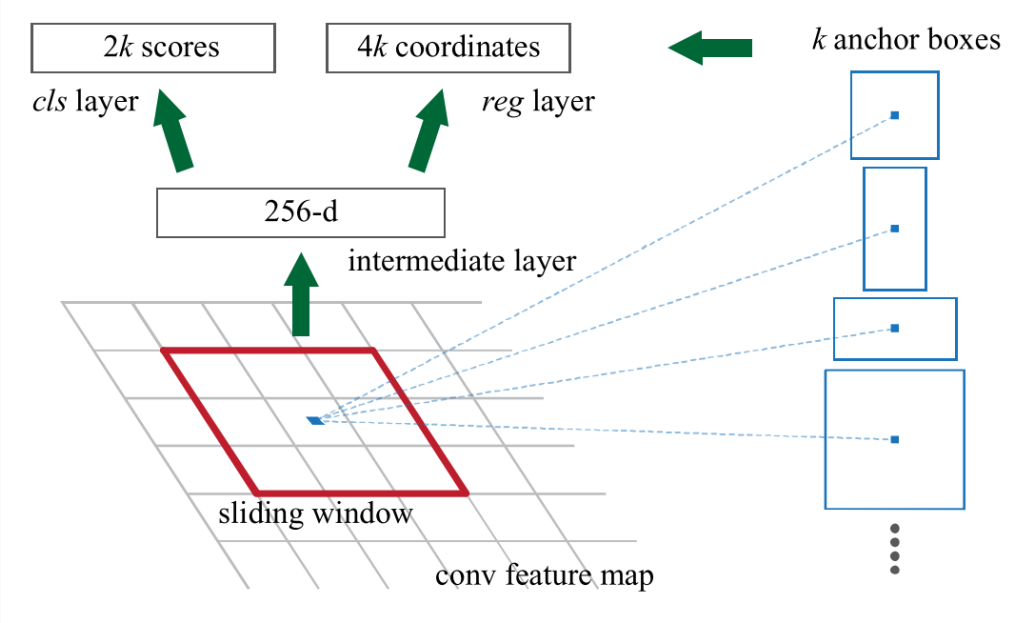


- Faster R-CNN aims to be an end-to-end detection pipeline.
- Regions are proposed by a **Region Proposal Network (RPN)**
- Region proposal and classification share the same convolutional layers.

S. Ren, K. He, et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Adv. in Neural Information Processing Systems (NIPS), 2015.

Faster R-CNN framework

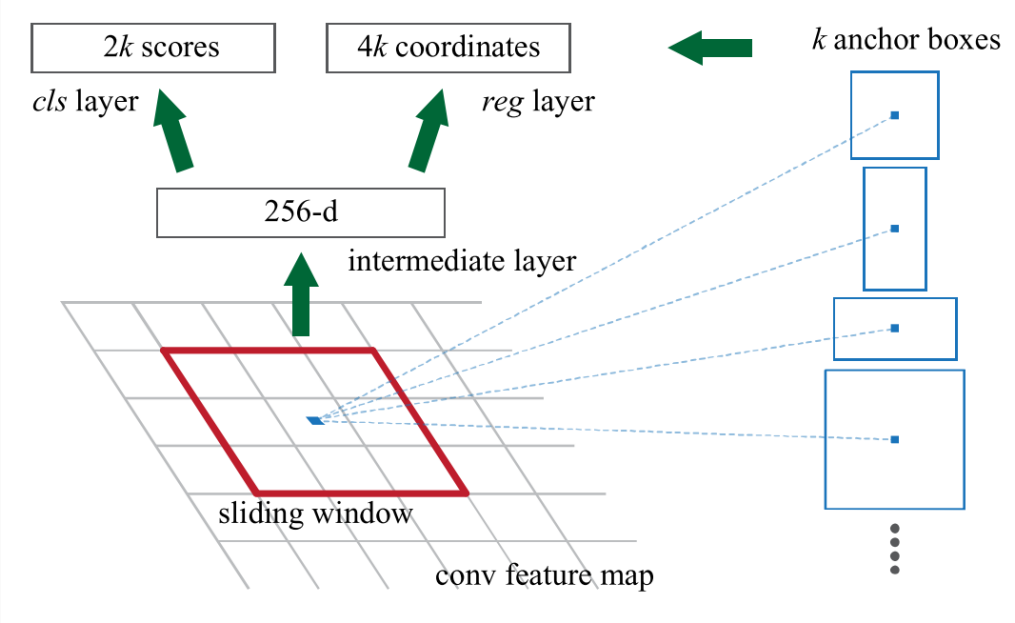
- RPN proposals are relative to fixed anchors.



S. Ren, K. He, et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Adv. in Neural Information Processing Systems (NIPS), 2015.

Faster R-CNN framework

- RPN proposals are relative to fixed anchors.
- Bounding boxes coordinates are refined in the final classification.



S. Ren, K. He, et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Adv. in Neural Information Processing Systems (NIPS), 2015.

Faster R-CNN in driving environments

Parameter optimization

- 1239 x 374 images scaled by a factor of ~1.33x (to height 500).

Faster R-CNN in driving environments

Parameter optimization

- 1239 x 374 images scaled by a factor of ~1.33x (to height 500).
- *DontCare* areas properly discarded.

Faster R-CNN in driving environments

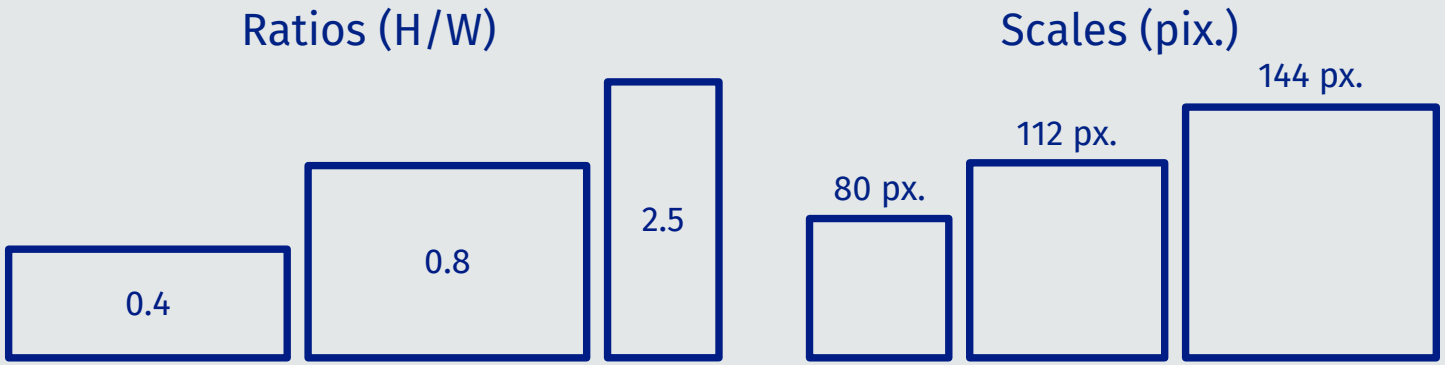
Parameter optimization

- 1239 x 374 images scaled by a factor of ~1.33x (to height 500).
- *DontCare* areas properly discarded.
- Hard, moderate and easy samples used to train.

Faster R-CNN in driving environments

Parameter optimization

- 1239 x 374 images scaled by a factor of ~1.33x (to height 500).
- *DontCare* areas properly discarded.
- Hard, moderate and easy samples used to train.
- RPN **anchors** selected to fit the objects in the environment.



Faster R-CNN in driving environments

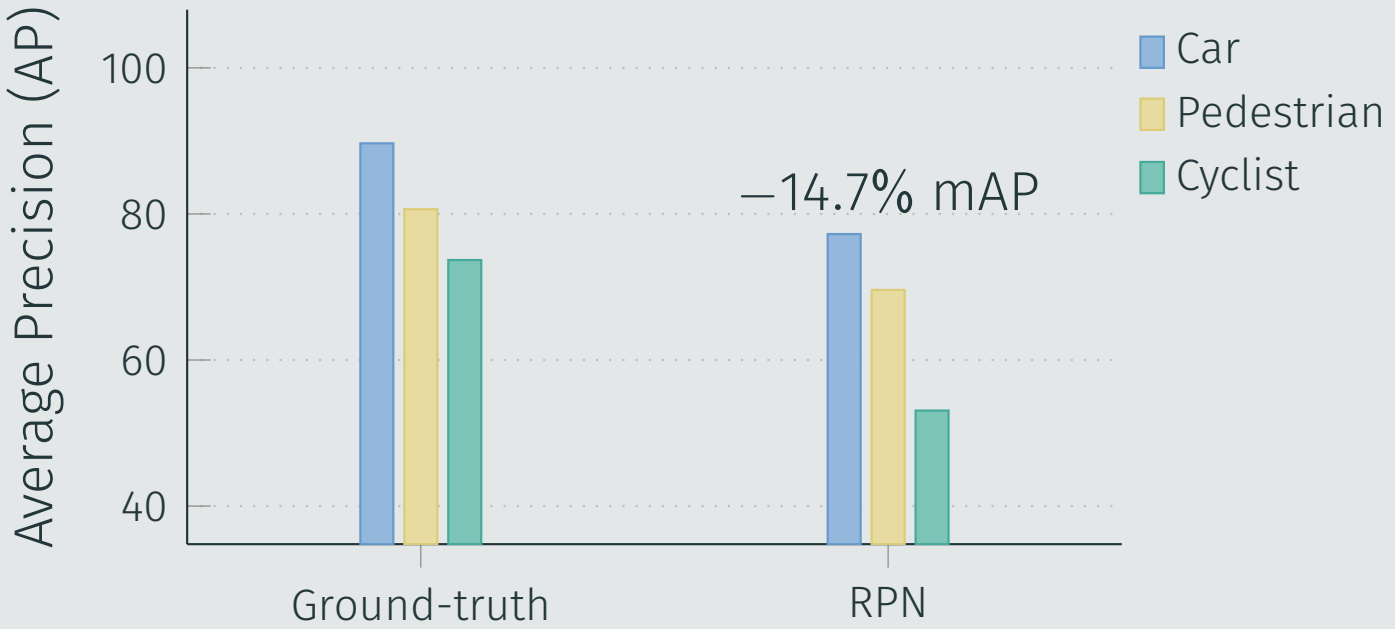
Limitations

- Accuracy highly dependent on the scale.

Faster R-CNN in driving environments

Limitations

- Accuracy highly dependent on the scale.
- Perfect proposals vs RPN proposals:



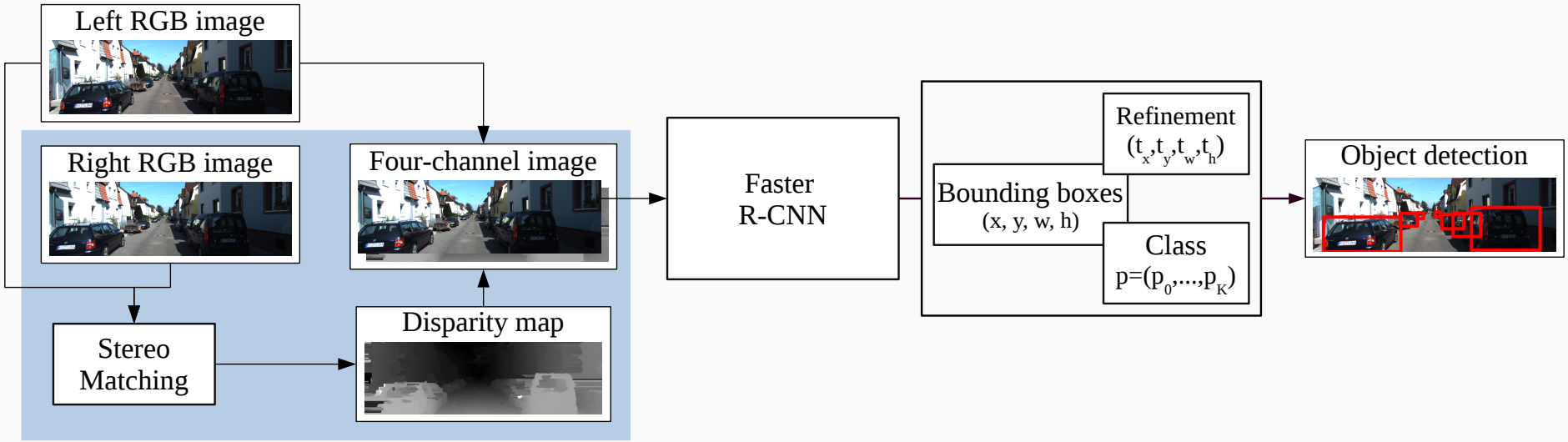
$$mAP = \frac{\sum_{i=0}^{N} AP_i}{N}$$

Proposed approach

System overview

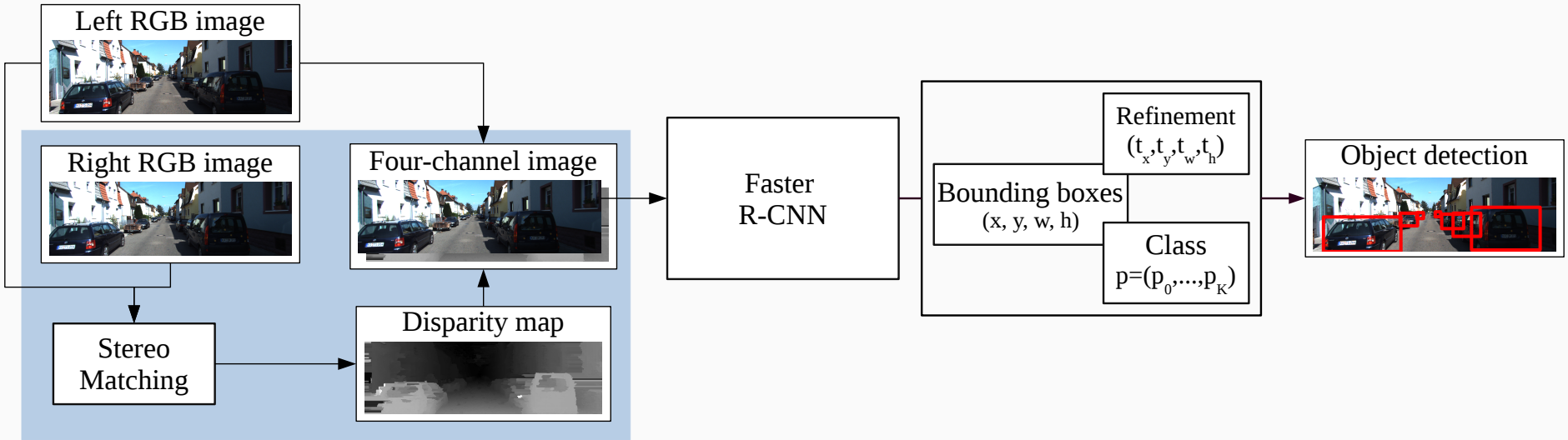
Goal:

- To use the frequently available stereo information in a simple, straightforward way to improve the detection accuracy.



System overview

- To use the frequently available stereo information in a simple, straightforward way to improve the detection accuracy.
- Intuitively, spatial reasoning provided by stereo may help to better segment the objects.

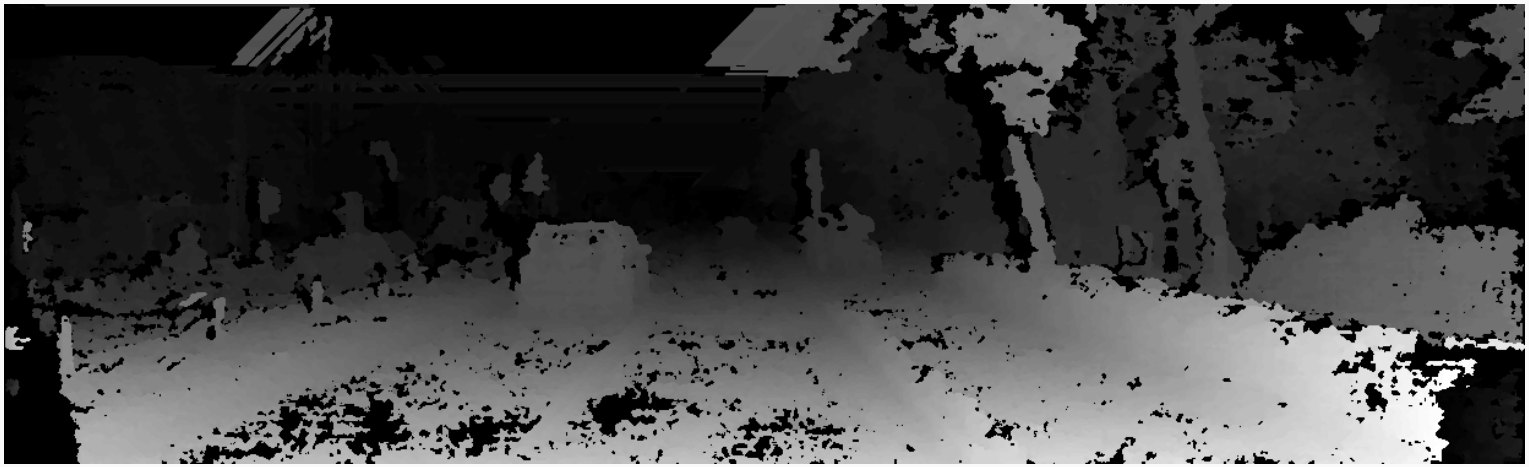


Stereo matching

- SG(B)M stereo-matching algorithm, is used.

Stereo matching

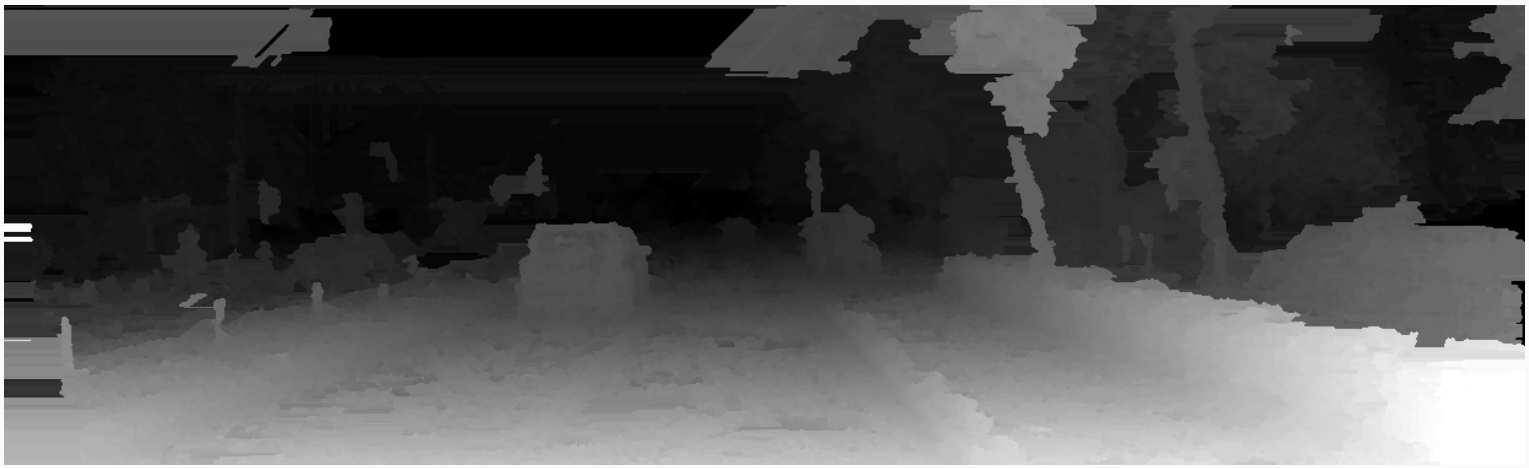
- SG(B)M stereo-matching algorithm, is used.
- Non-defined pixels may introduce spurious errors.



Stereo matching

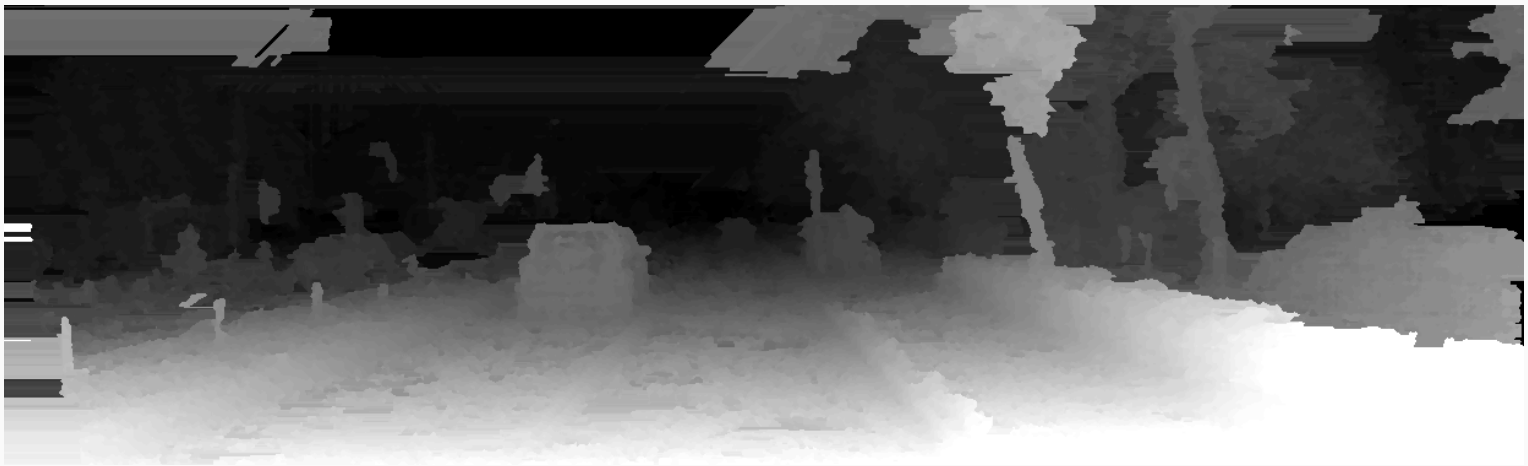
- SG(B)M stereo-matching algorithm, is used.
- Non-defined pixels may introduce spurious errors.
- Thus, a simple background interpolation is applied.

$$i\_disp[u_{NON\_DEFINED}, v] = min(disp[u_{LAST\_DEFINED}, v], disp[u_{NEXT\_DEFINED}, v])$$



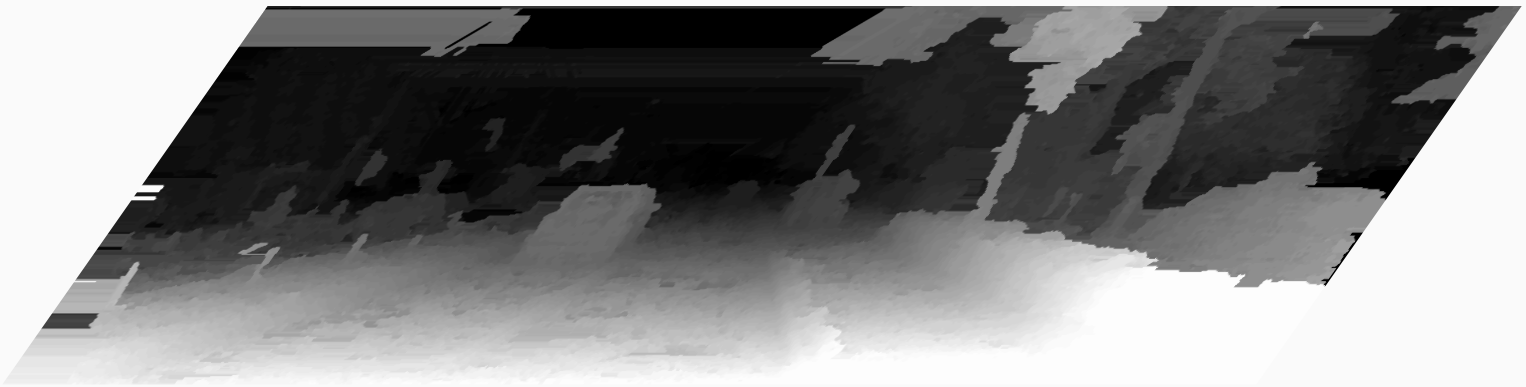
CNN input

- The disparity map is normalized before entering the CNN: values between 0-63 (6 meters and further) are scaled to 0-255. Disparities between 64-127 are clipped to 255.



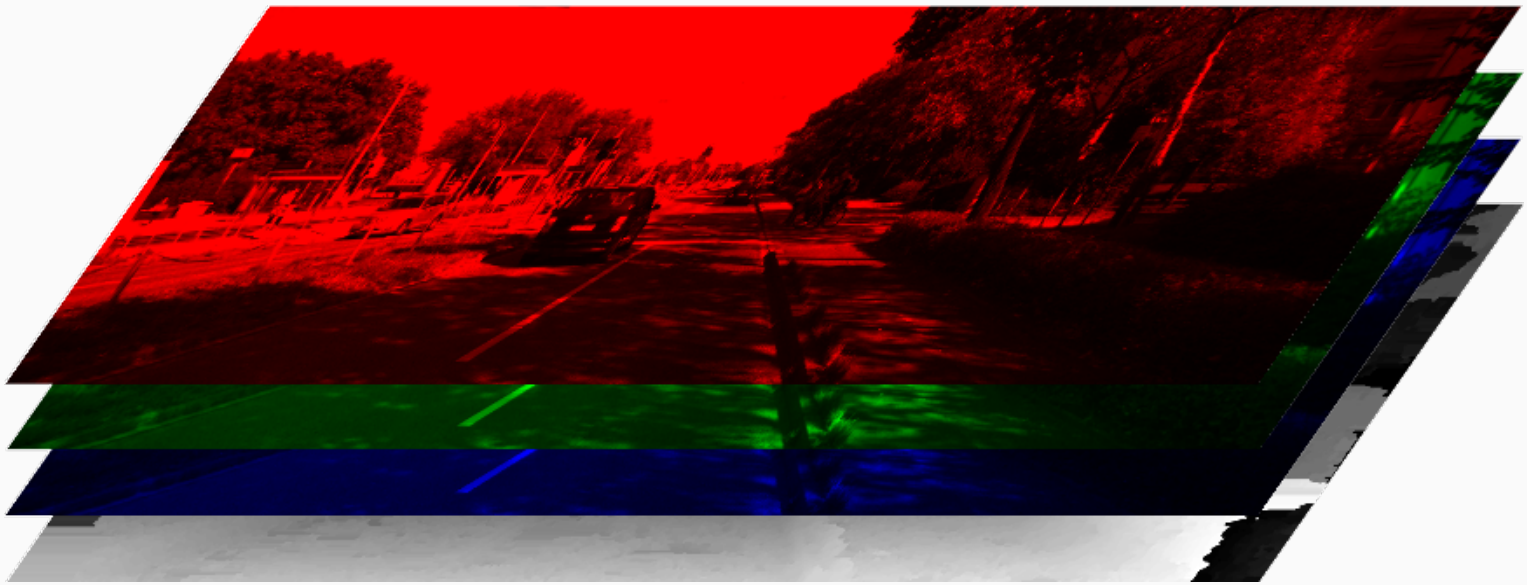
CNN input

- The disparity map is normalized before entering the CNN: values between 0-63 (6 meters and further) are scaled to 0-255. Disparities between 64-127 are clipped to 255.
- The normalized disparity map is appended to the RGB image as a **fourth channel**.



CNN input

- The disparity map is normalized before entering the CNN: values between 0-63 (6 meters and further) are scaled to 0-255. Disparities between 64-127 are clipped to 255.
- The normalized disparity map is appended to the RGB image as a **fourth channel**.



CNN architecture and initial weights

- No major changes are required in the CNN architecture itself.

CNN architecture and initial weights

- No major changes are required in the CNN architecture itself.
- 4-channel filters in the first convolutional layer.

CNN architecture and initial weights

- No major changes are required in the CNN architecture itself.
- 4-channel filters in the first convolutional layer.

Initial values

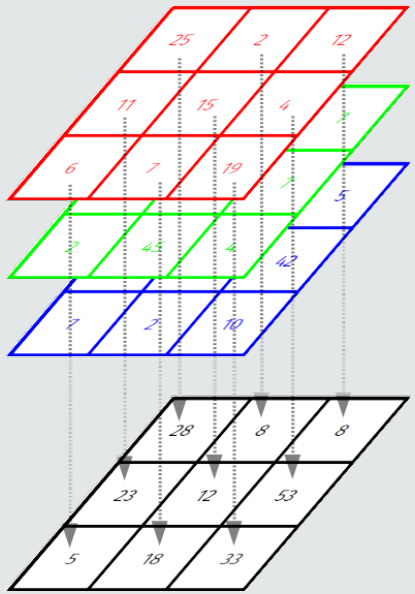
- The shared convolutional layers are initialized using a pre-trained model for ImageNet classification, as is standard practice.

CNN architecture and initial weights

- No major changes are required in the CNN architecture itself.
- 4-channel filters in the first convolutional layer.

Initial values

- The shared convolutional layers are initialized using a pre-trained model for ImageNet classification, as is standard practice.
- In order to re-use the available Faster R-CNN models, weights for the fourth channel are computed as the mean of the RGB weights.



Results

Setup

Architecture

- VGG 16-layer architecture was used in the experiments.
- With the smaller Zeiler-Fergus architecture, the accuracy drops.

Setup

Architecture

- VGG 16-layer architecture was used in the experiments.
- With the smaller Zeiler-Fergus architecture, the accuracy drops.

Training set

- KITTI training set was divided into two splits.
- **Training** is performed over 7 categories but **tests** are limited to 3 categories: *Car*, *Pedestrian* and *Cyclist*.

Setup

Architecture

- VGG 16-layer architecture was used in the experiments.
- With the smaller Zeiler-Fergus architecture, the accuracy drops.

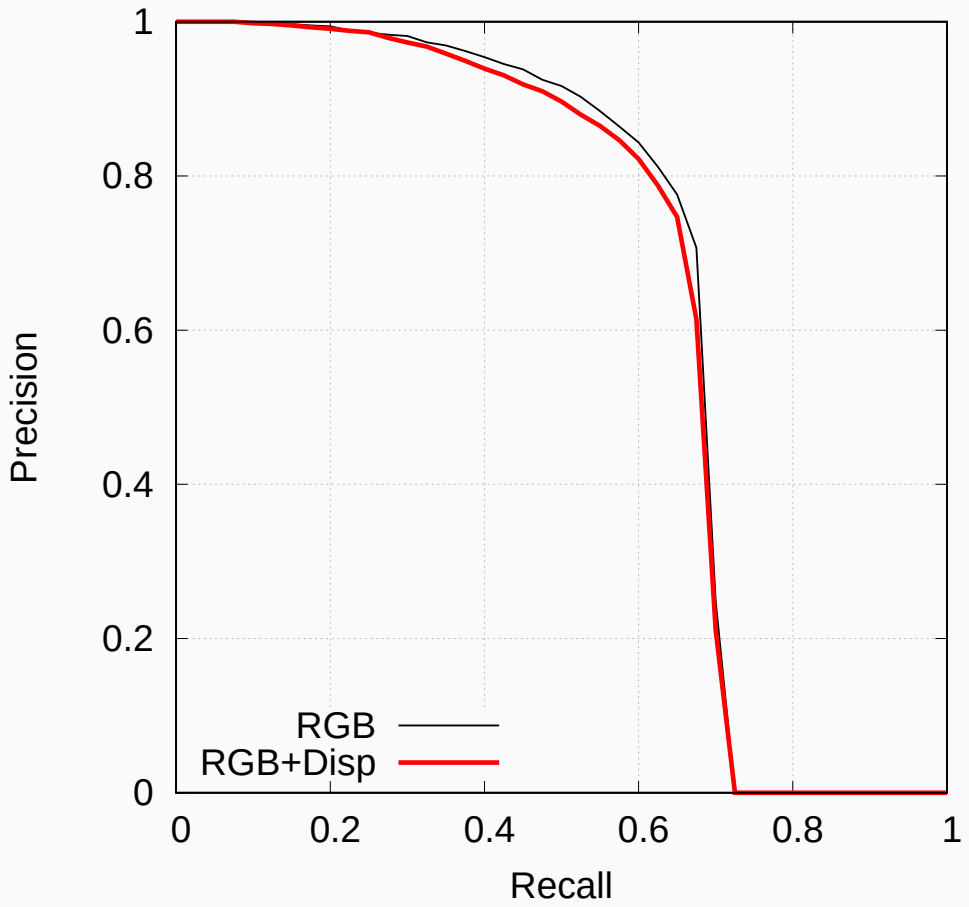
Training set

- KITTI training set was divided into two splits.
- **Training** is performed over 7 categories but **tests** are limited to 3 categories: *Car*, *Pedestrian* and *Cyclist*.

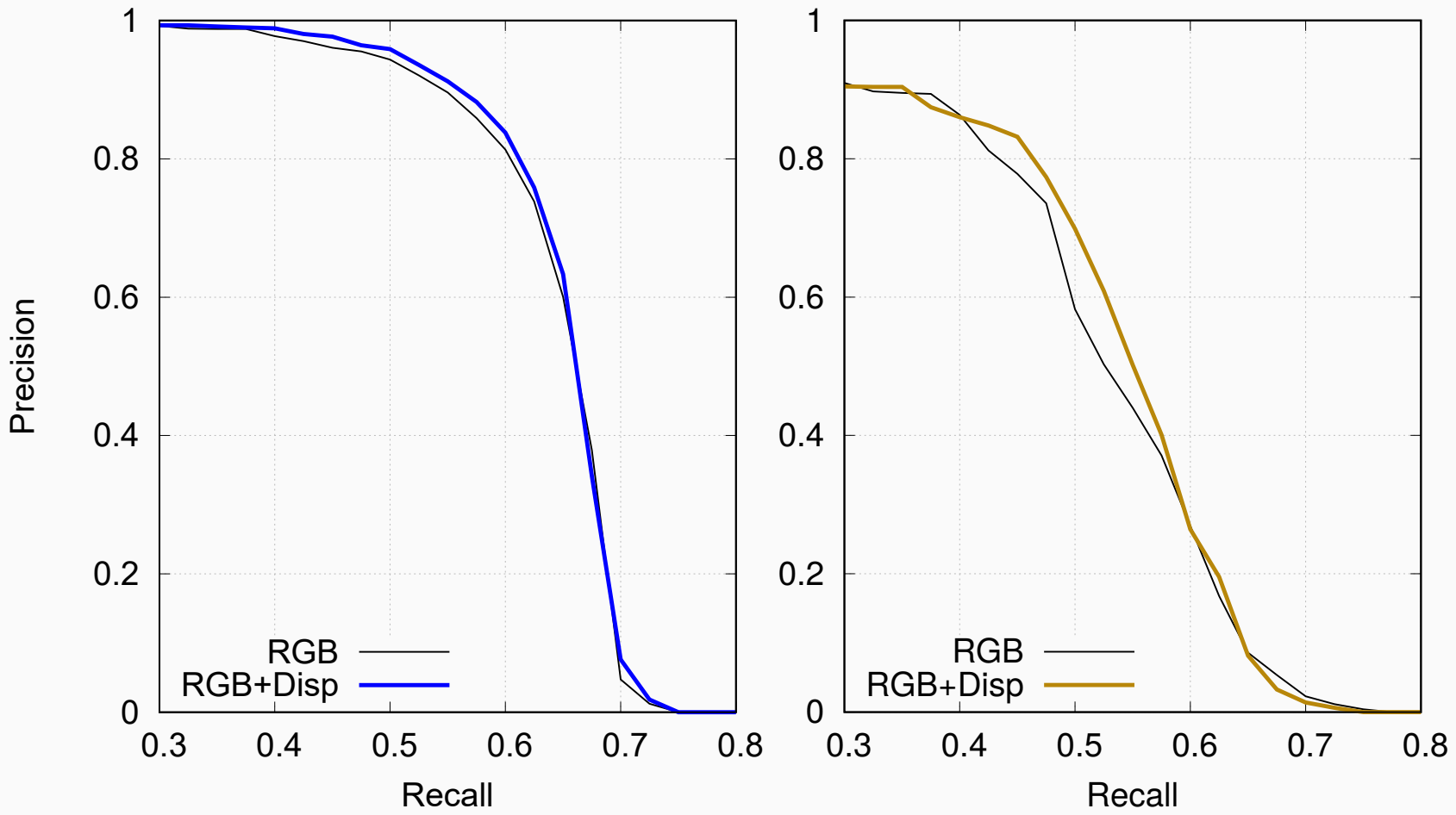
Parameters

80k iterations, learning rate: 0.001 (x0.1 after 50k)

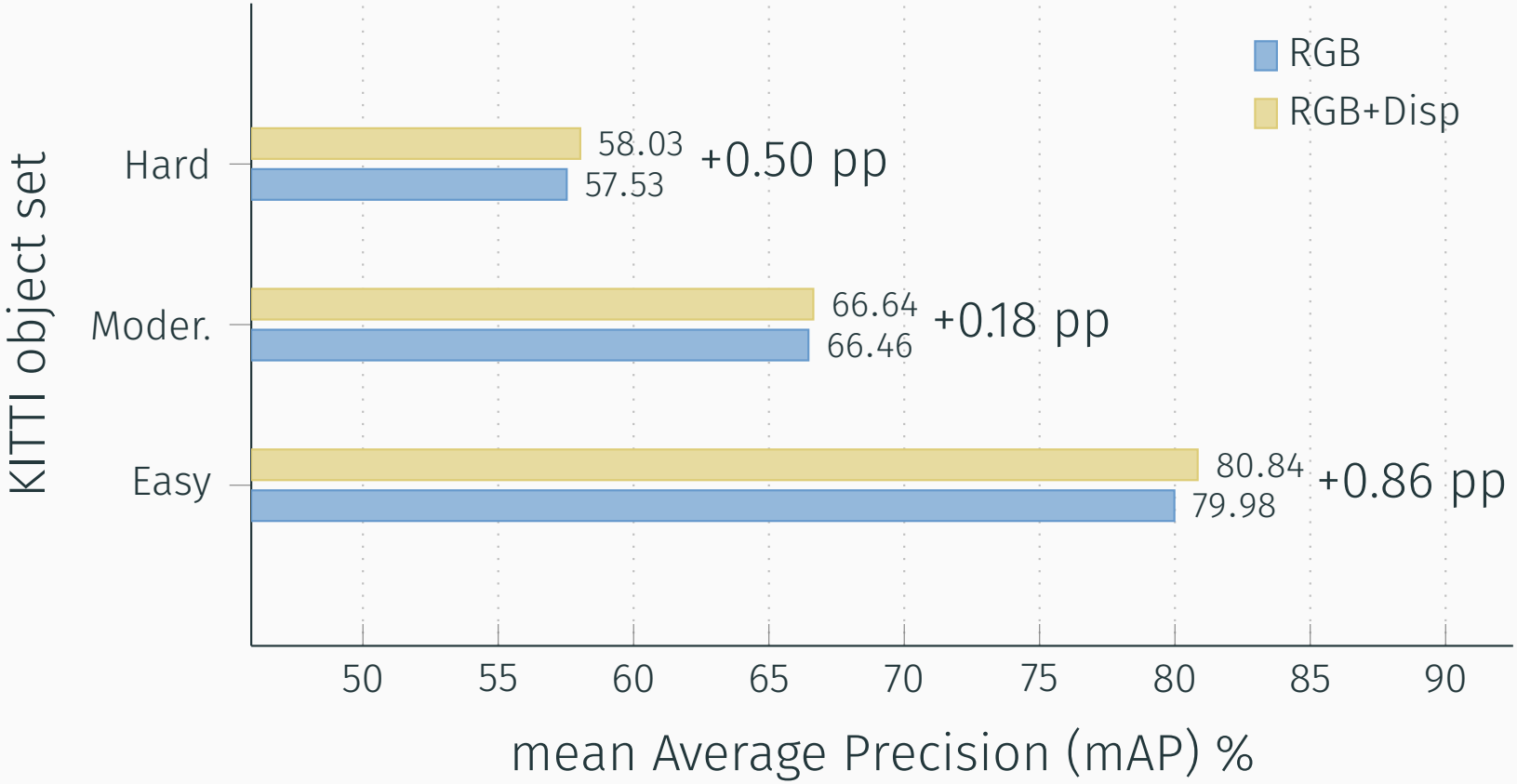
Precision-recall curves (Car)



Precision-recall curves (Pedestrian & Cyclist)



Detection AP (%)



	Car	Pedestrian	Cyclist
AP (moderate)	76.03%	69.18%	54.72%

IVVI 2.0 platform · Intelligent Systems Laboratory

Intelligent Vehicle based on Visual Information

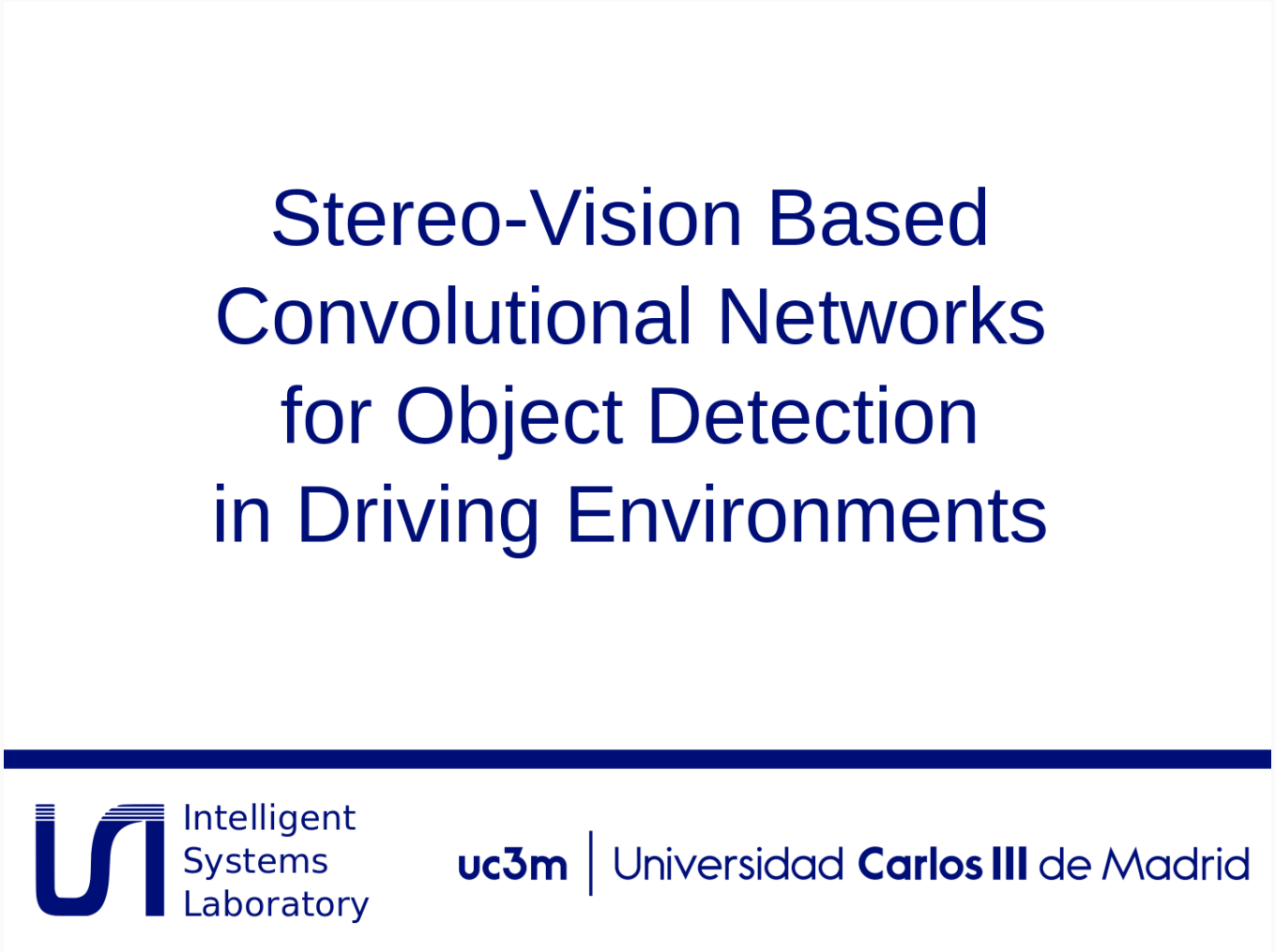
- Platform to test Advanced Driver Assistance Systems.

IVVI 2.0 platform · Intelligent Systems Laboratory

Intelligent Vehicle based on Visual Information

- Platform to test Advanced Driver Assistance Systems.
- Sensors: stereo camera, LiDAR, IMU, GPS, ...

Video


Stereo-Vision Based
Convolutional Networks
for Object Detection
in Driving Environments
Intelligent
Systems
Laboratory
uc3m | Universidad Carlos III de Madrid

Conclusions

Conclusions

- Convolutional Neural Networks represent a robust paradigm for object detection in driving environments.

Conclusions

- Convolutional Neural Networks represent a robust paradigm for object detection in driving environments.
- We have proven that **stereo information** can be used to enhance the accuracy of CNNs.

Conclusions

- Convolutional Neural Networks represent a robust paradigm for object detection in driving environments.
- We have proven that **stereo information** can be used to enhance the accuracy of CNNs.

Future experiments

- CNN architectures with better accuracy/computation time trade-off (ResNets).
- Several stereo matching algorithms.
- Different hole-filling techniques (white noise).
- Methods to use features from early layers.

Thank you

Intelligent Systems Laboratory

uc3m

Stereo-Vision Based Convolutional Networks for Object Detection in Driving Environments

Carlos Guindel (cguindel@ing.uc3m.es) , David Martín, José M. Armingol

Sixteenth International Conference on Computer Aided Systems Theory
Intelligent Transportation Systems and Smart Mobility Workshop

Intelligent Systems Laboratory · Universidad Carlos III de Madrid