Analysis of the Influence of Training Data on Road User Detection

Carlos Guindel, David Martín, José María Armingol, and Christoph Stiller

Intelligent Systems Lab

uc3m | Universidad Carlos III de Madrid

mrt Institut für Mess- und Regelungstechnik

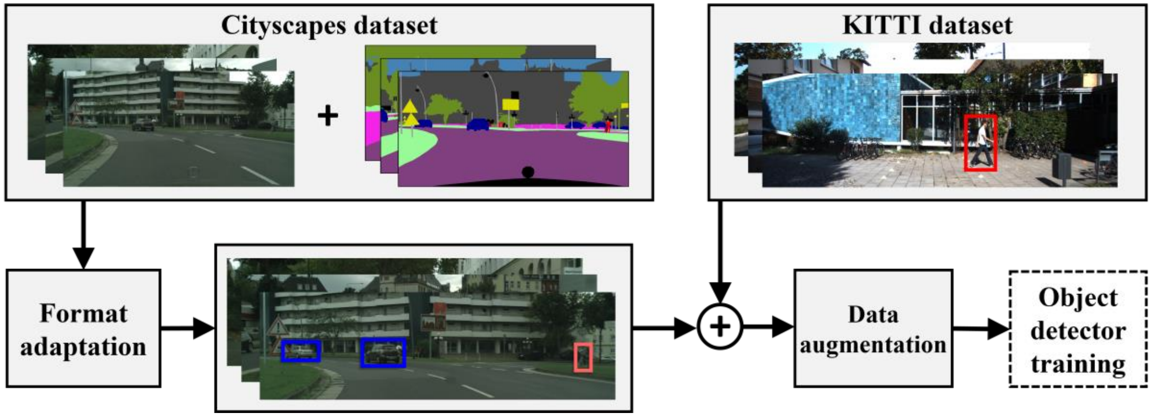
KIT Karlsruher Institut für Technologie

ICVES 2018

20th IEEE International Conference on Vehicular Electronics and Safety
Madrid · 12 September 2018

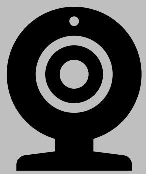
Agenda

- Motivation and goals
- Experimental setup
- Analysis
- Conclusion



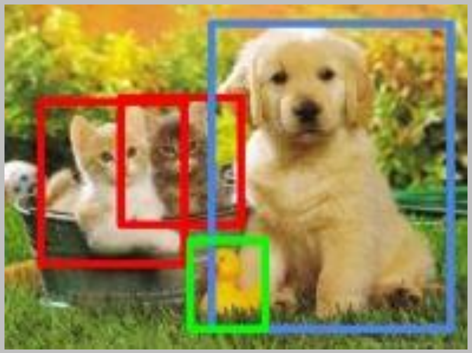
Motivation

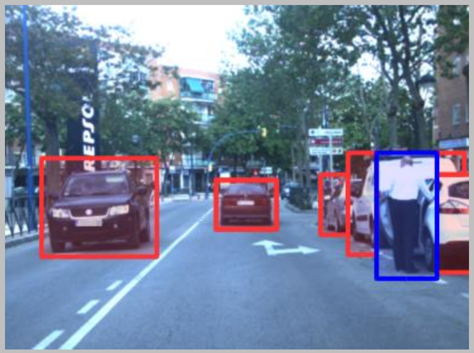
Object detection



Computer vision

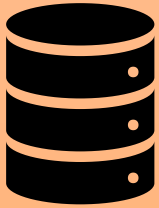
Autonomous driving





Deep Learning

Data



Instance segmentation (e.g., Mask R-CNN)

Motivation

Object detection

Computer vision

Autonomous driving

Deep Learning

IMAGENET

450 000+ images
200 categories

COCO
Common Objects in Context

200 000+ images
80 categories

…

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

7 481 images
9 categories

CITYSCAPES DATASET

2 975 images
10 categories

Data

Motivation

Object detection

Autonomous driving

Deep Learning

Different labels

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

7 481 images
9 categories

CITYSCAPES DATASET

2 975 images
10 categories

Data

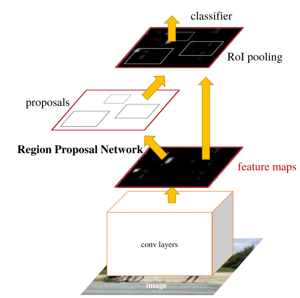
Goals

Research is often narrow-focused on the development of new architectures and models.
R-FCN, SSD; ResNet, Inception, MobileNet,...

Instead, we investigate:

The improvement provided by introducing additional samples into the training process.

The possibility of using heterogeneous labels in a multi-task learning method.



Faster R-CNN
State-of-the-art object detection meta-architecture

Datasets

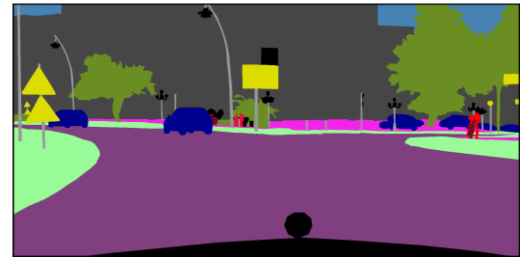
The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

Training

3,712 images

Validation

3,769 images



CITYSCAPES DATASET

Training

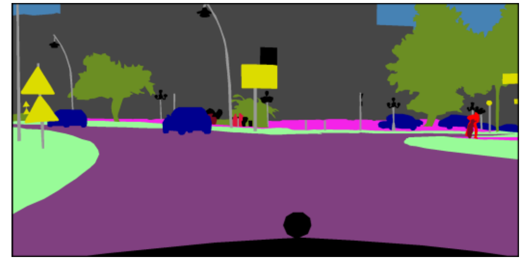
2,975 images

category	KITTI			Cityscapes
	train	val	total	total
Car	10 753	10 963	21 716	21 637
Pedestrian	2 104	2 172	4 276	15 788
Cyclist	594	600	1 194	1 481

X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3D Object Proposals for Accurate Object Class Detection," in Advances in Neural Information Processing Systems (NIPS), 2015, pp. 424–432.

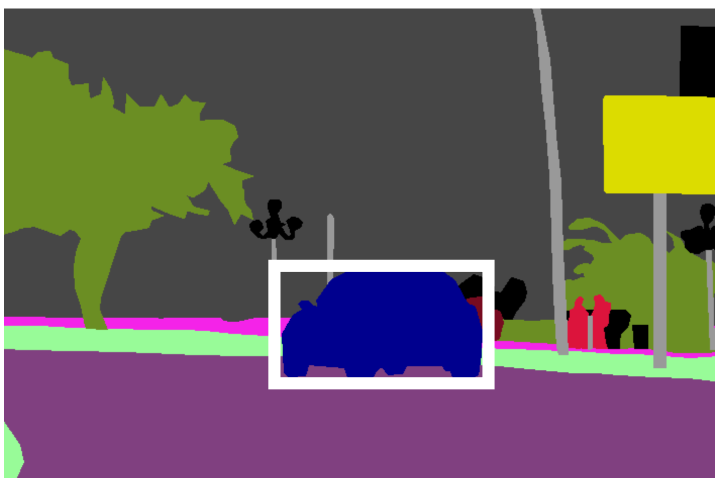
Adaptation

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago



CITYSCAPES DATASET

① Semantic labeling to bounding boxes



Instantiable objects

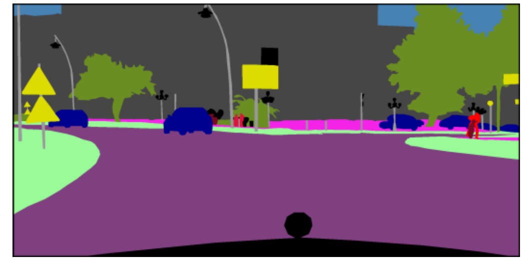
<u>Minimum enclosing box</u>

Categories:

- Person → Pedestrian
- Rider + Bicycle → Cyclist

Adaptation

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago



CITYSCAPES DATASET

② Occlusion & truncation

Occlusion

Cityscapes labels contain foreground-background ordering

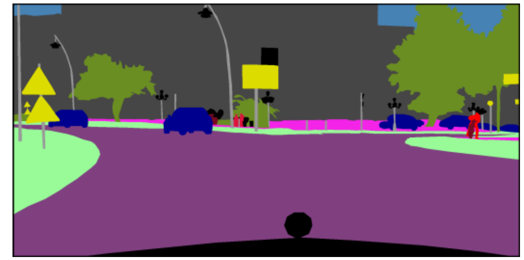
$$\text{Degree of occlusion} = \frac{\text{intersection}}{\text{background}}$$

Truncation

Whenever any of the sides of the b. box coincides with the image boundaries

Adaptation

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago



CITYSCAPES DATASET

③ Resolution/FOV

2048 × 1024

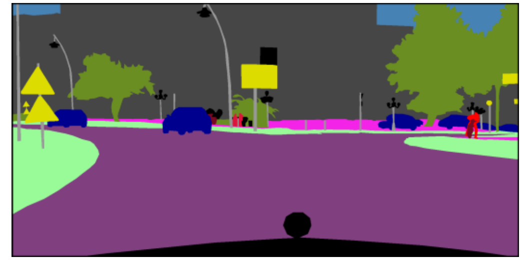
(KITTI: 1224 × 370 aprox.)

2048 × 620

Removes the hood and the Mercedes-Benz emblem

Adaptation

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago



CITYSCAPES DATASET

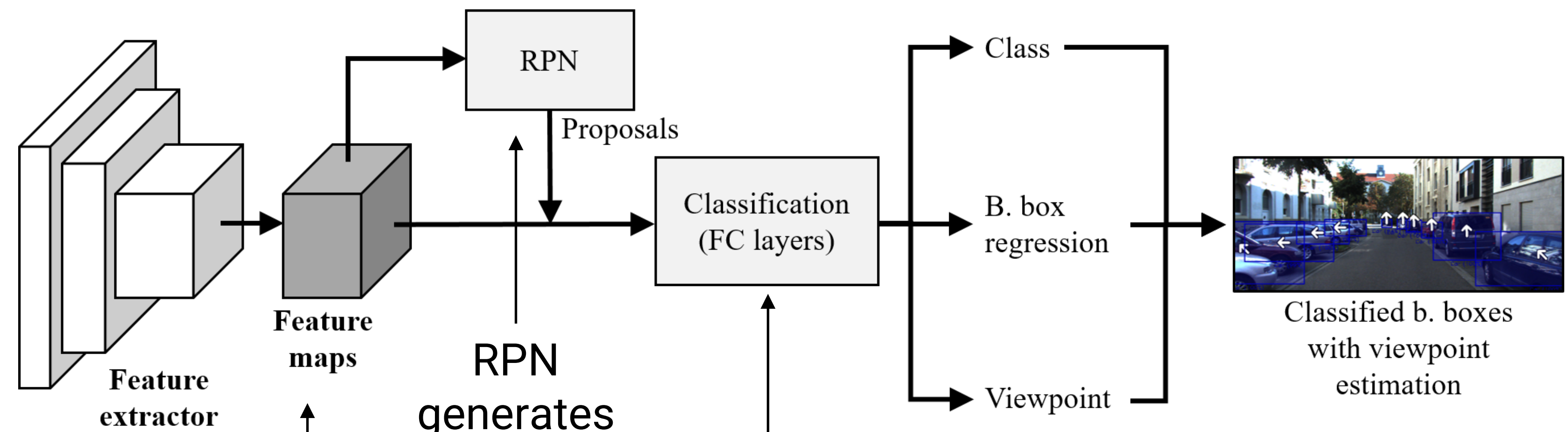
④ Difficulty levels

We ignore samples not meeting the KITTI's ***Hard*** level requirements:

- Larger than **25 pixels**
- Max **occlusion**: "Difficult to see": level 2 (KITTI) or 75% (Cityscapes)
- Maximum **truncation**: 50% (KITTI) or no truncation (Cityscapes)

Object detection method

Vanilla **Faster R-CNN** | Modified



RPN generates proposals

RGB input

Network *backbone* **VGG-16**

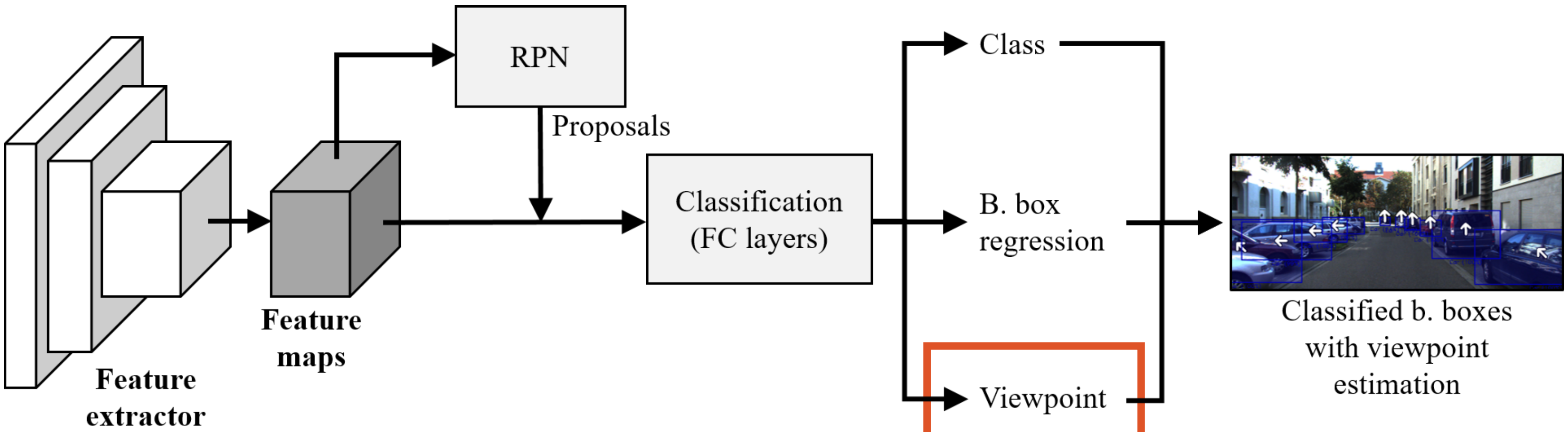
Features **shared** among all the tasks

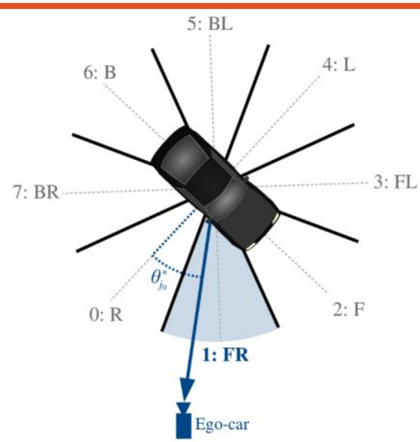
Several FC layers stacked

Each proposal is classified, refined *and given an estimated observation angle*

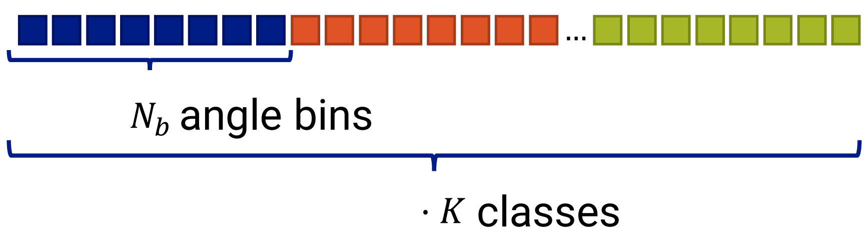
C. Guindel, D. Martin, and J. M. Armingol, "Fast Joint Object Detection and Viewpoint Estimation for Traffic Scene Understanding," accepted for publication in: Intelligent Transportation Systems Magazine.

Object detection method





$$r^k = (r_0^k, \ldots, r_{N_b}^k) \text{ for } k = 0, \ldots, K$$



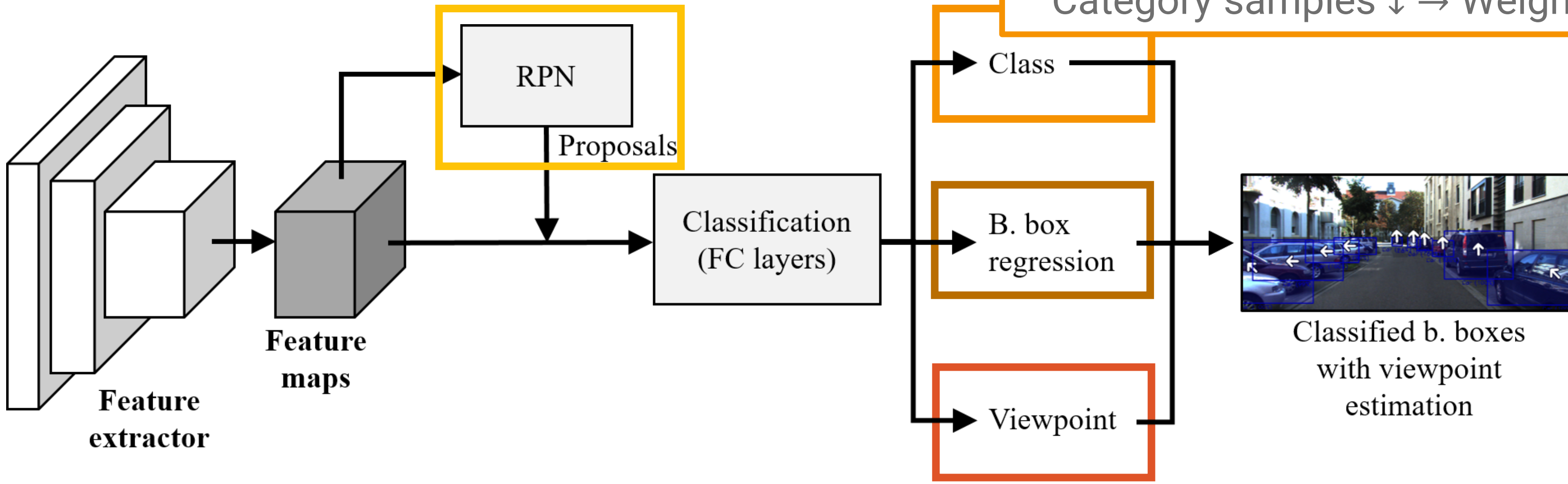
$N_b \cdot K$ elements

Multinomial logistic loss $\longrightarrow$ $\frac{1}{N_{B_2}} \sum_{i \in B_2} [v_i \geq 1]\, L_{cls}(\mathbf{r}_i^{v_i}, w_i)$

C. Guindel, D. Martin, and J. M. Armingol, “Fast Joint Object Detection and Viewpoint Estimation for Traffic Scene Understanding,” accepted for publication in: Intelligent Transportation Systems Magazine.

Multi-task loss

Weighted multinomial logistic loss
Category samples ↓ → Weight ↑

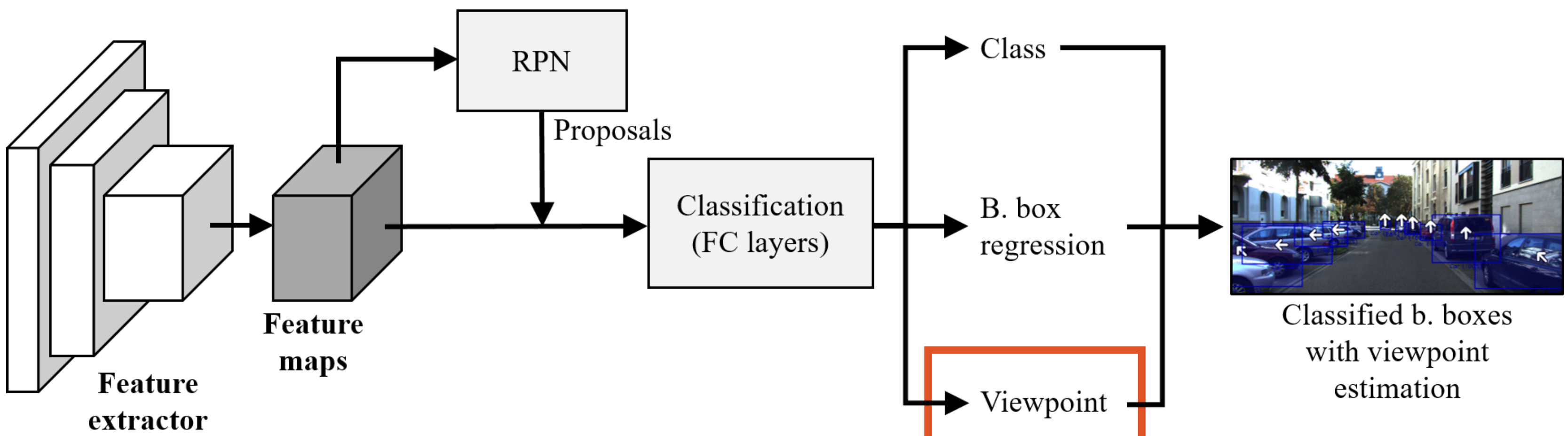


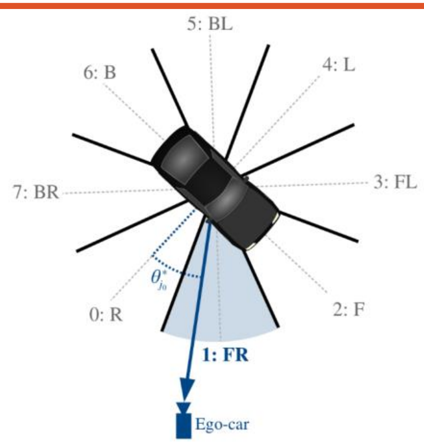
Loss = **Objectness + Proposal regr.** + **Class** + **B.box regr.** + ***Viewpoint***

Multinomial logistic loss → $\frac{1}{N_{B_2}} \sum_{i \in B_2} [v_i \geq 1] \, L_{cls}(\mathbf{r}_i^{v_i}, w_i)$

C. Guindel, D. Martin, and J. M. Armingol, “Fast Joint Object Detection and Viewpoint Estimation for Traffic Scene Understanding,” accepted for publication in: Intelligent Transportation Systems Magazine.

Multi-task loss





The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

Viewpoint loss $\neq$ 0

CITYSCAPES DATASET

No viewpoint/orientation annotations

Viewpoint loss = 0

Multinomial logistic loss $\longrightarrow$ $$\frac{1}{N_{B_2}} \sum_{i \in B_2} [v_i \geq 1] \, L_{cls}(\mathbf{r}_i^{v_i}, w_i)$$

Assessment method

Evaluation metrics

Precision Recall

Average precision (AP)

$$AP = \frac{1}{N}\sum_r \tilde{p}(r) \qquad \tilde{p}(r) = \max_{\tilde{r}:\tilde{r}>r} p(\tilde{r})$$

Assess object detection

Average orientation similarity (AOS)

$$AOS = \frac{1}{N}\sum_r \tilde{s}(r) \qquad \tilde{s}(r) = \max_{\tilde{r}:\tilde{r}>r} s(\tilde{r})$$

Assess object detection
AND orientation

Cosine similarity Recall

Common Testbed

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

Validation set

✓ 3,769 images

Training Parameters

- 1- image batch
- SGD, initial lr = 0.001
- Step decay schedule
  0.1× every 50k iterations
- 80k iterations

Experiment 1: Combined datasets

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

Training

3,712 images

Add training samples

Adapted

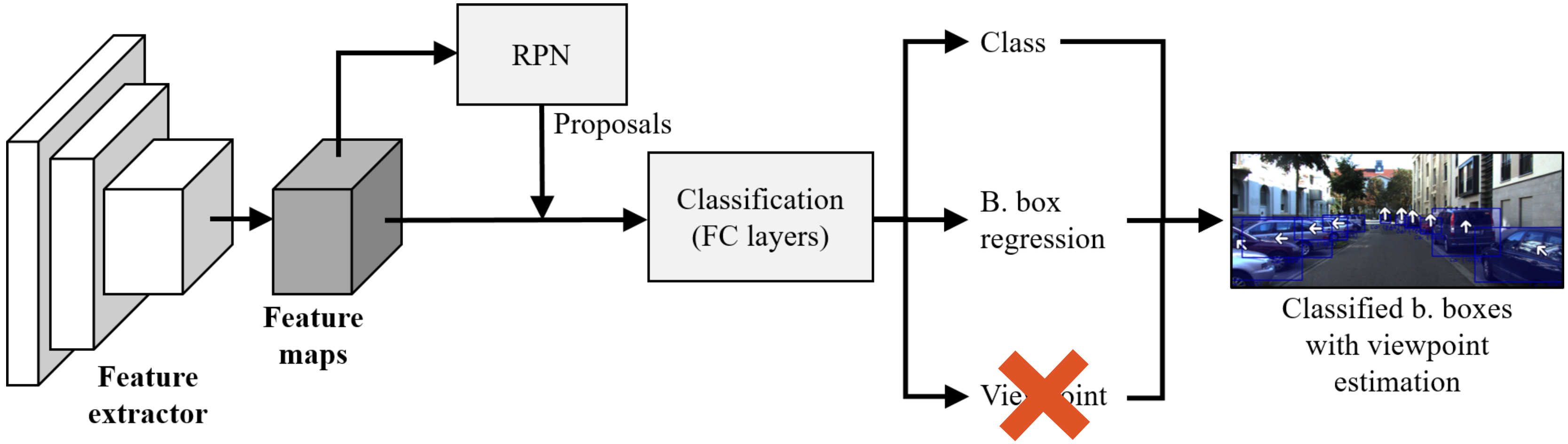
CITYSCAPES DATASET

Training

2,975 images

- In each training iteration, an image (single batch) is randomly chosen from a mix of both datasets.
- Tests without (a) and with (b) viewpoint estimation branch

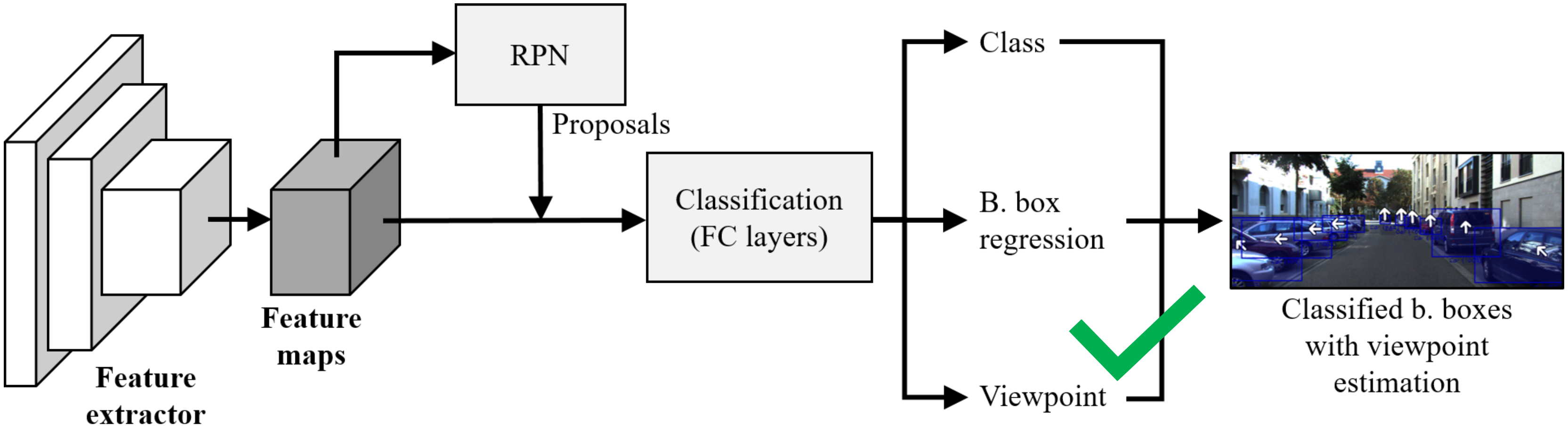
Experiment 1: Combined datasets (a)



Detection

category	tr. data	Easy	Mod.	Hard	
Car	KITTI	90.05	79.32	70.04	
	Cityscapes	81.37	63.66	53.47	
	KITTI + CS	**90.31**	**84.94**	**70.33**	**+5.62 AP**
Pedestrian	KITTI	75.80	67.17	58.58	
	Cityscapes	72.00	63.92	55.33	
	KITTI + CS	**77.77**	**68.72**	**60.05**	**+1.55 AP**
Cyclist	KITTI	77.47	56.96	54.64	
	Cityscapes	63.09	50.14	46.85	
	KITTI + CS	**82.90**	**62.50**	**58.05**	**+5.54 AP**

Experiment 1: Combined datasets (b)



Detection +
Orientation

Two alternative strategies:

1 Pick images from a **KITTI+Cityscapes** mix

Viewpoint = 0 when a Cityscapes sample is chosen

2 Pre-train with **Cityscapes**, fine-tune with **KITTI**

Train without viewpoint branch and transfer weights to the complete model

Experiment 1: Combined datasets (b)

Detection + Orientation

1. Pick images from a **KITTI+Cityscapes** mix
2. Pre-train with **Cityscapes**, fine-tune with **KITTI**

category	tr. data	Detection (AP) Easy	Detection (AP) Mod.	Detection (AP) Hard	Orientation (AOS) Easy	Orientation (AOS) Mod.	Orientation (AOS) Hard
Car	KITTI	90.01	79.03	69.67	88.26	77.35	67.97
	1 KITTI + CS	**90.39**	84.59	70.21	**88.68**	82.79	68.57
	2 KITTI (w. CS pret.)	90.33	**86.16**	**70.58**	88.63	**84.43**	**69.01**
Pedestrian	KITTI	71.19	64.05	55.75	65.31	57.62	50.01
	1 KITTI + CS	**76.32**	**67.98**	**59.11**	**67.83**	**59.65**	**51.69**
	2 KITTI (w. CS pret.)	74.54	66.01	57.68	67.33	59.01	51.52
Cyclist	KITTI	77.33	54.87	52.89	69.73	48.79	47.06
	1 KITTI + CS	**86.11**	**68.49**	**63.46**	**77.66**	**61.23**	**56.83**
	2 KITTI (w. CS pret.)	83.18	60.37	57.35	75.55	54.36	51.74

+5.6 AP
+1.6 AOS

+3.9 AP
+2 AOS

+13.6 AP
+12.4 AOS

Experiment 2: Can we get rid of ImageNet?

~~Pre-training~~ → The KITTI Vision Benchmark Suite

Pre-training with ImageNet (generalist dataset) generates good initial weights

CITYSCAPES DATASET + The KITTI Vision Benchmark Suite ?

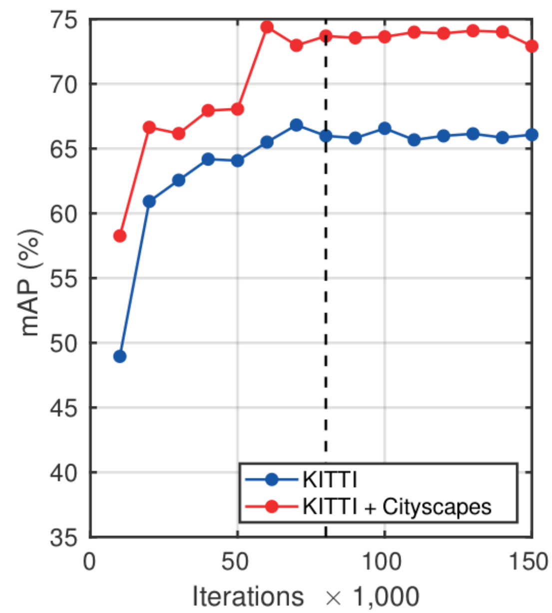
init.	tr. data	Detection (mAP)			Orientation (mAOS)		
		Easy	Mod.	Hard	Easy	Mod.	Hard
Yes	KITTI	**79.51**	**65.98**	**59.44**	**74.43**	**61.25**	**55.02**
No	K. + CS	53.80	42.99	37.25	47.93	39.13	33.00

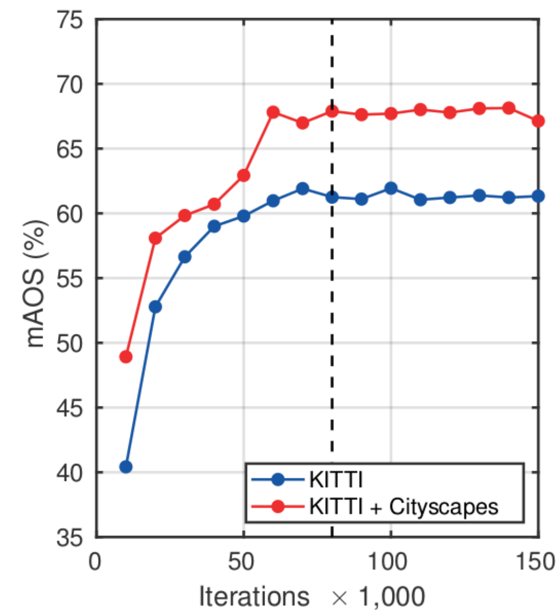
−22.99 mAP **−22.12 mAOS**

✗ Initialization with a large dataset is still an essential requierement to achieve a proper generalization ability

Experiment 3: Overfitting

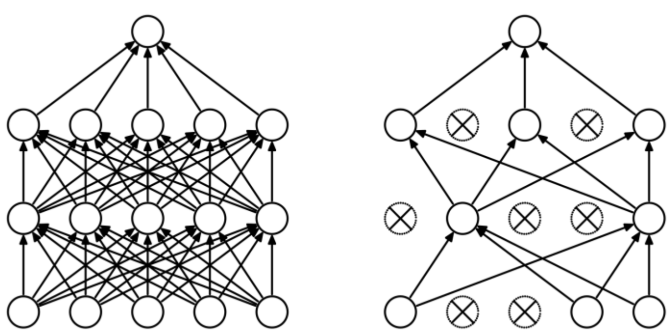
Performance on the validation set vs # of iterations





No symptoms of overfitting

Dropout? $p = 0.5$



dropout	Detection (mAP)			Orientation (mAOS)		
	Easy	Mod.	Hard	Easy	Mod.	Hard
No	**79.51**	**65.98**	**59.44**	**74.43**	**61.25**	**55.02**
Yes	79.20	65.34	58.43	73.77	60.73	54.16

−0.64 mAP **−0.52 mAOS**

No apparent benefit

Experiment 4: Missing labels

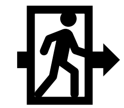
The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

+

CITYSCAPES DATASET

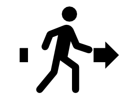
Observation angle annotations ✓

No observation angle annotations ✗



Detection + Orientation

+6.64 mAOS



Orientation

?

Additional measure

Mean precision in pose estimation (MPPE)

- Discrete viewpoint estimation is a classification problem
- MPPE is the mean of the elements on the main diagonal of the confusion matrix (correctly predicted viewpoint bins)

	Frontal	Frontal–Left	Left	Rear–Left	Rear	Rear–Right	Right	Frontal–Right
Frontal	.69	.06		.03	.22			.01
Frontal–Left	.01	.99						
Left	.01	.16	.69	.07			.04	.02
Rear–Left	.01			.82	.06			.11
Rear	.11	.01	.05	.02	.77	.04		
Rear–Right	.01	.25			.04	.68	.02	
Right		.03	.02	.01	.01	.10	.69	.15
Frontal–Right	.14			.29				.57

R. J. López-Sastre, T. Tuytelaars, and S. Savarese, "Deformable part models revisited: A performance evaluation for object category pose estimation," in Proc. IEEE International Conference on Computer Vision (ICCV), 2011, pp. 1052–1059.

Experiment 4: Missing labels

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

\+ CITYSCAPES DATASET

Observation angle annotations ✓

No observation angle annotations ✗

Detection + Orientation	Orientation
+6.64 mAOS	**+1,52 MPPE**

category	tr. data	Easy	Mod.	Hard
Car	KITTI	**92.24**	80.93	69.29
	KITTI + CS	92.13	**83.40**	**71.72**
Pedestrian	KITTI	**59.03**	51.02	43.71
	KITTI + CS	57.74	**51.35**	**44.41**
Cyclist	KITTI	**70.71**	49.84	49.00
	KITTI + CS	64.95	**51.60**	**49.11**

+2.47 MPPE (Car)

+0.33 MPPE (Pedestrian)

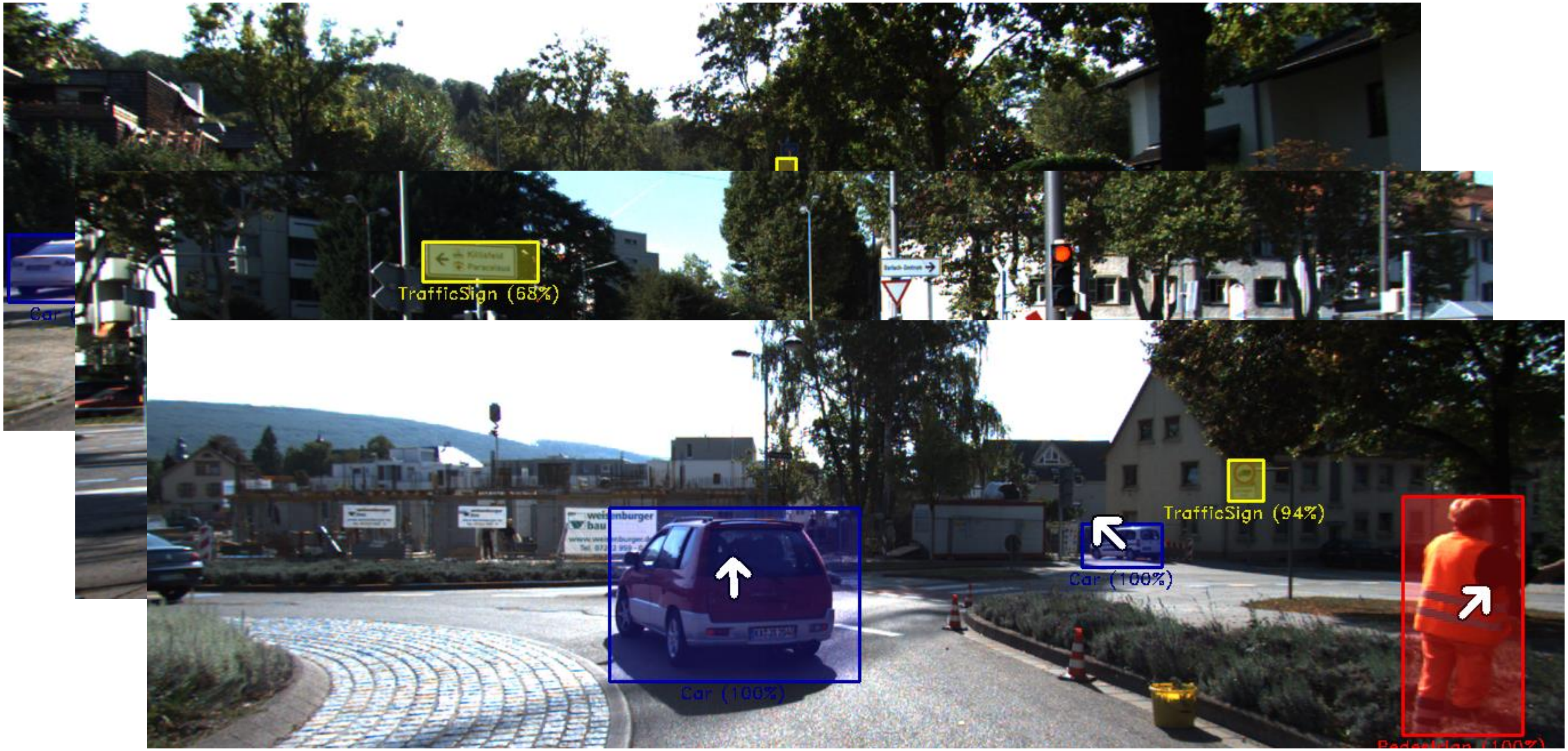
+1.76 MPPE (Cyclist)

Missing labels does not hurt orientation estimation performance ✓

Experiment 5: Mixed labels

Including all categories

Car, Truck, Pedestrian, Cyclist, Train, Traffic Sign → **Cityscapes-only**



Experiment 6: Data augmentation

Horizontal flip + texture augmentations

https://github.com/aleju/**imgaug**

Choose a random subset of them, between 0 and 4

Add

$[-40, 40]$

Multiplication

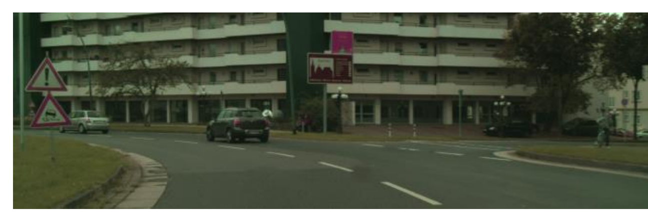
$[0.5, 1.5]$

Gaussian noise

$\mathcal{N}(0, 5.1^2)$

Saturation

$[-20, 20]$ (H, S)



Experiment 6: Data augmentation

Two separated experiments

1 Only **KITTI**

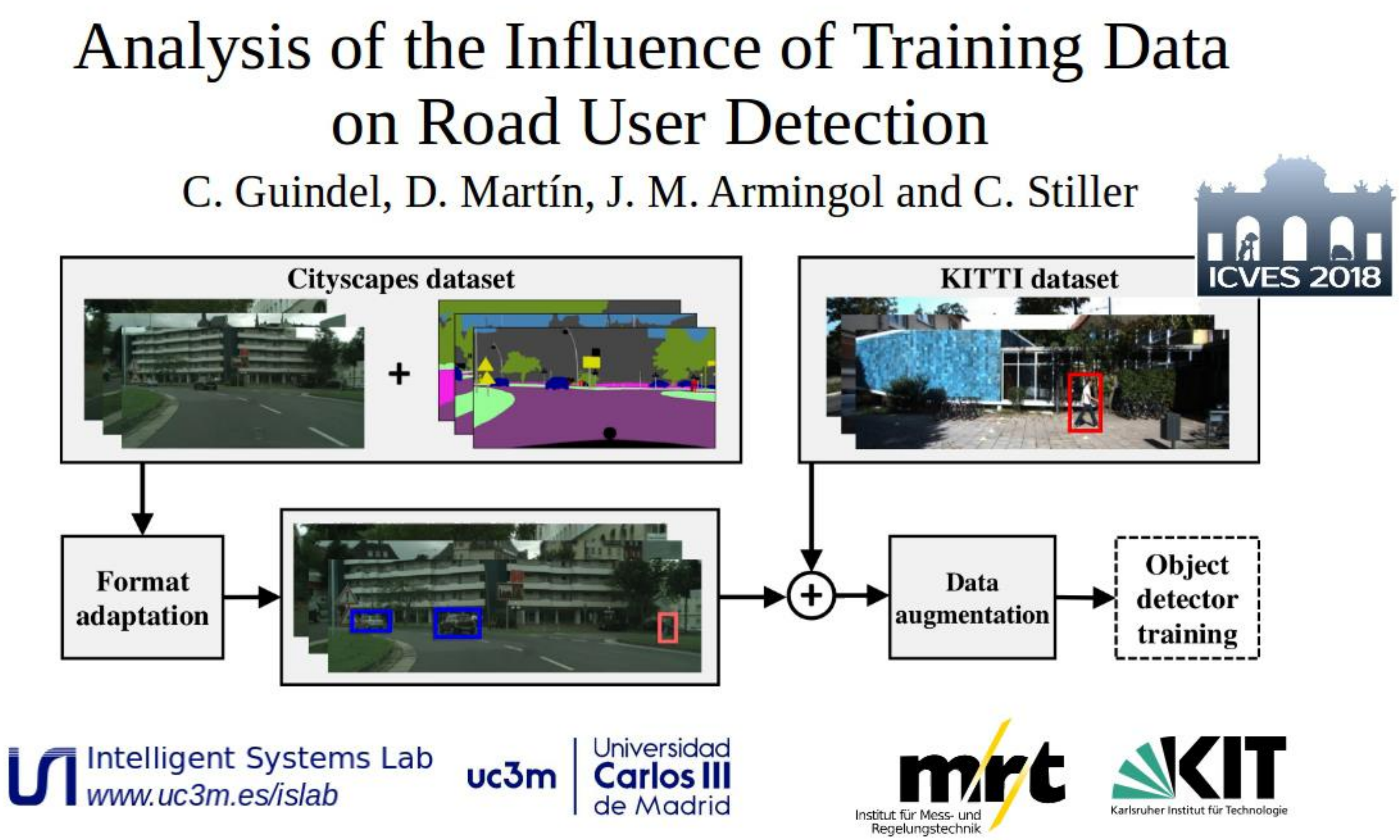
To assess the overall effect of the augmentation techniques

2 **Cityscapes + KITTI**

Augmentation could help mitigate the difference between both sets of images

	tr. data	aug.	Detection (mAP)			Orientation (mAOS)			
			Easy	Mod.	Hard	Easy	Mod.	Hard	
1	K.	No	79.51	**65.98**	**59.44**	74.43	**61.25**	**55.02**	✗ No apparent benefit
		Yes	**80.39**	65.87	58.96	**74.56**	61.00	54.38	
2	K. + CS	No	**84.27**	73.69	64.26	**78.06**	67.89	59.03	~ Limited benefit
		Yes	83.96	**74.14**	**65.16**	77.95	**68.09**	**59.59**	

Comparison


Analysis of the Influence of Training Data
on Road User Detection
C. Guindel, D. Martín, J. M. Armingol and C. Stiller
ICVES 2018
Cityscapes dataset
+
KITTI dataset
Format adaptation
Data augmentation
Object detector training
Intelligent Systems Lab
www.uc3m.es/islab
uc3m
Universidad Carlos III de Madrid
mrt
Institut für Mess- und Regelungstechnik
KIT
Karlsruher Institut für Technologie

Conclusion

Modestly enhancing the training data can lead to notable improvements on the results obtained by a CNN object detector

The variability introduced by Cityscapes samples can achieve a non-negligible improvement, even when evaluated on the KITTI dataset

Results pave the way for future works taking advantage of multiple data sources

Thank You

20th IEEE International Conference on Vehicular Electronics and Safety
Madrid · 12 September 2018