



UNIVERSITY OF PADUA
UNIVERSITA' DEGLI STUDI DI PADOVA

Instance Segmentation of Urban Street Scenes

Francesco Bari, Eleonora Signor

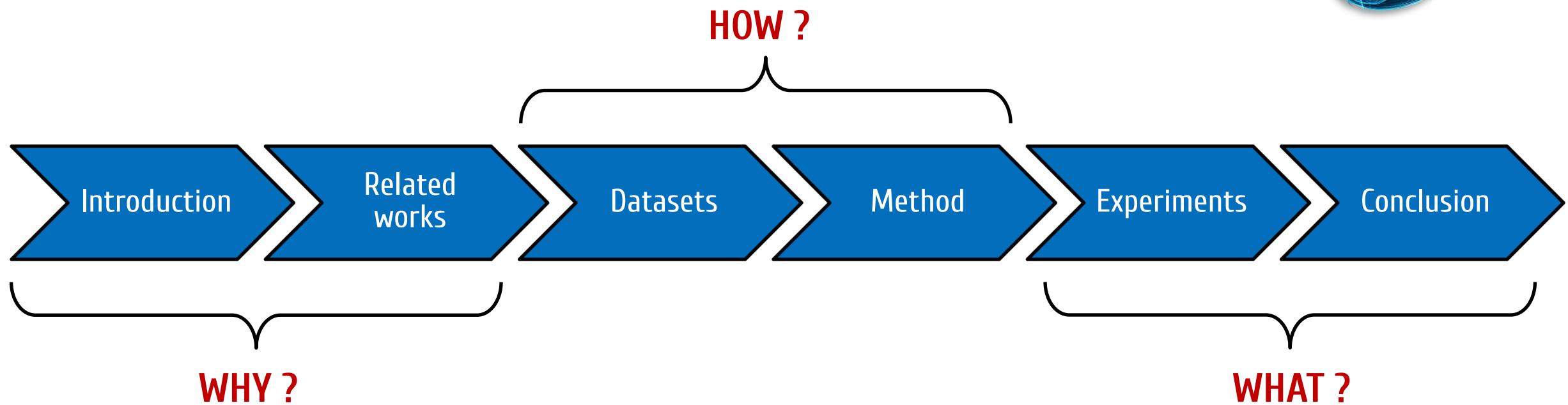
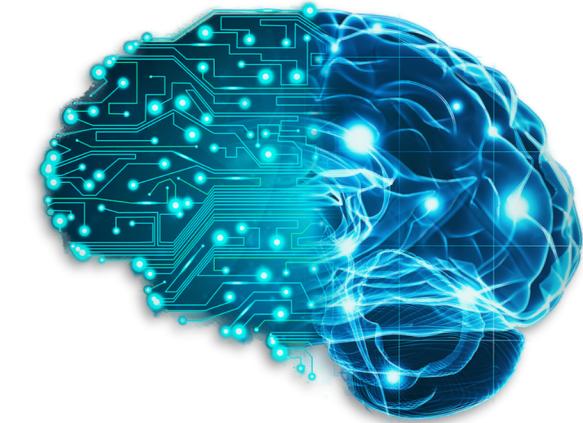
Vision and Cognitive Systems - Prof. Lamberto Ballan

21 Febbraio 2022



DIPARTIMENTO
MATEMATICA

➤ Table of contents



➤ Goal: → Instance Segmentation of Urban Street Scenes

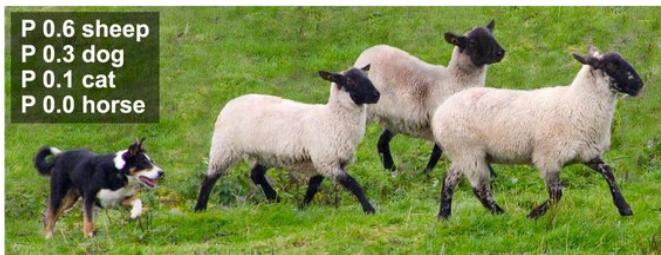
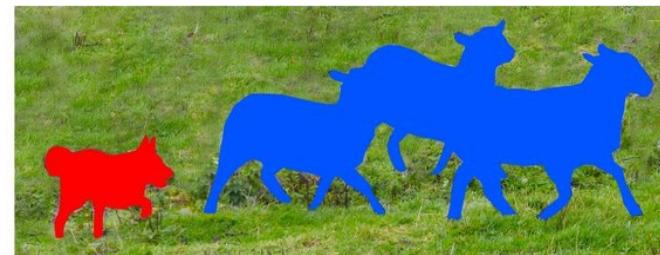
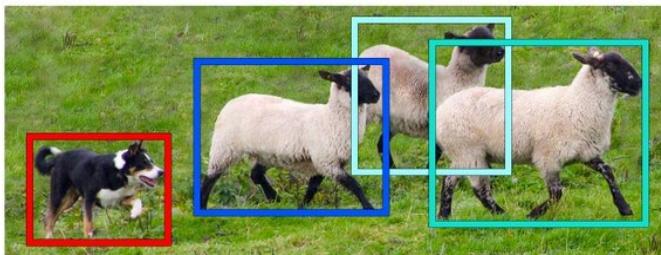


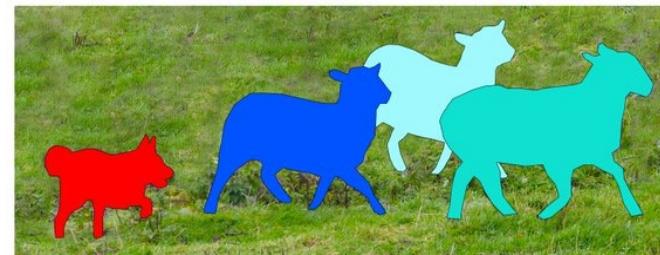
Image Recognition



Semantic Segmentation



Object Detection

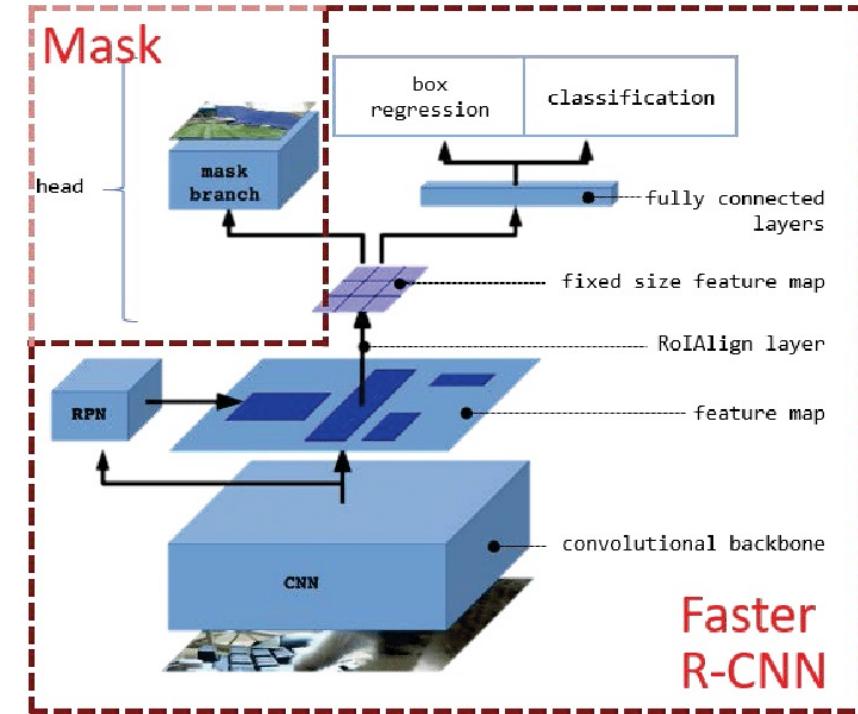
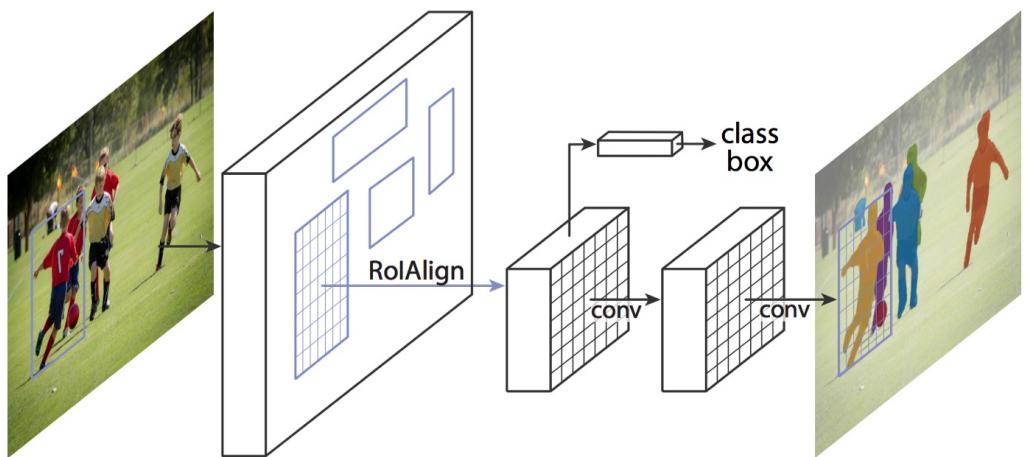


Instance Segmentation



➤ Mask R-CNN: Box-based two stage approach

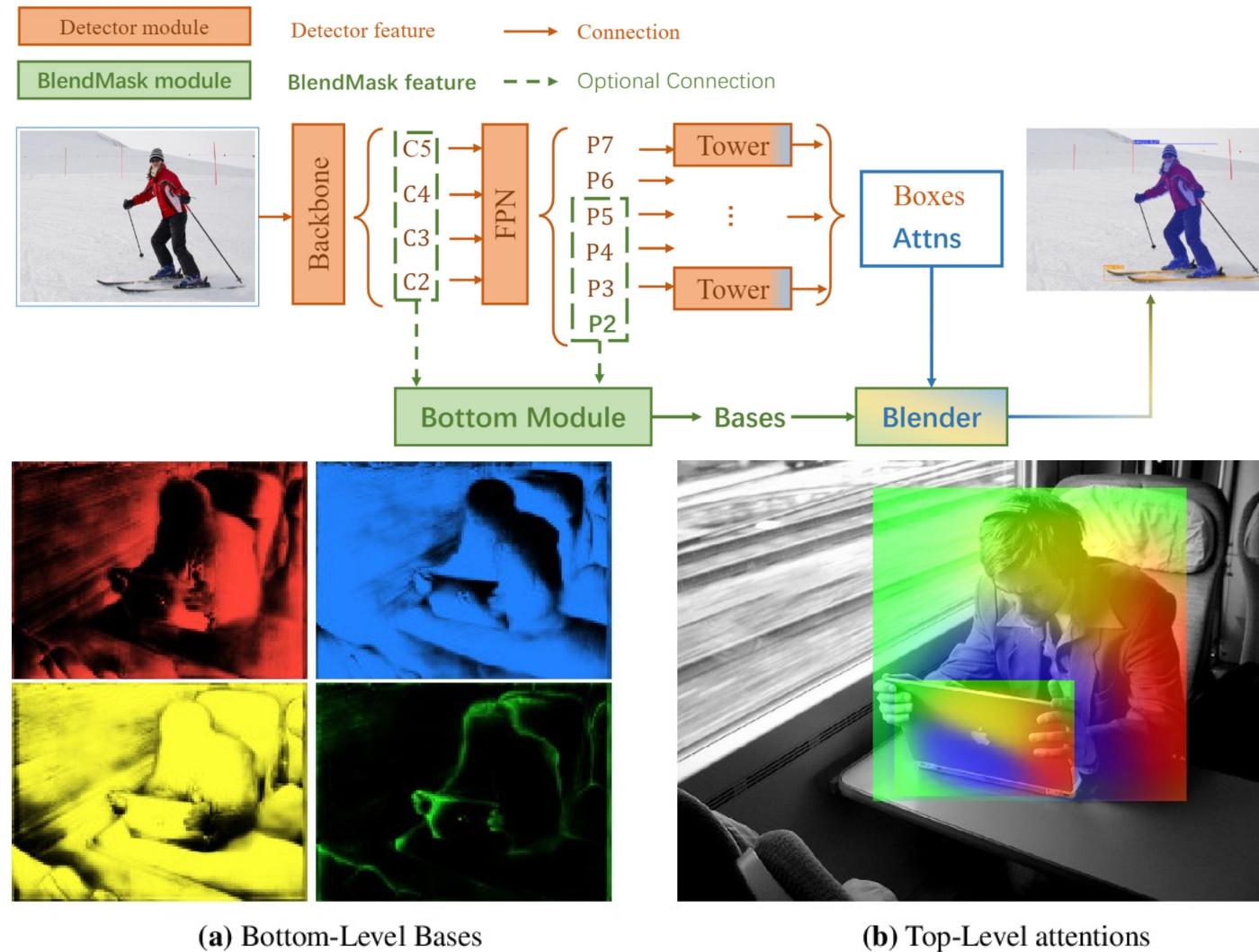
- Generates a segmentation mask for each instance
- Region Proposal Network
- Class label, bounding box offset and object mask
- Backbone and head
- RoIAlign



✖ COMPLICATED SCENARIOS

✖ MULTI-TASK NETWORK

➤ BlendMask: Anchor box-free one-stage approach



- **Bottom module:** deals with the production of the bases
- **Top layer:** a single convolution layer, one for each tower. Predictions attention instances
- **Blender module:** combines bases with attentions

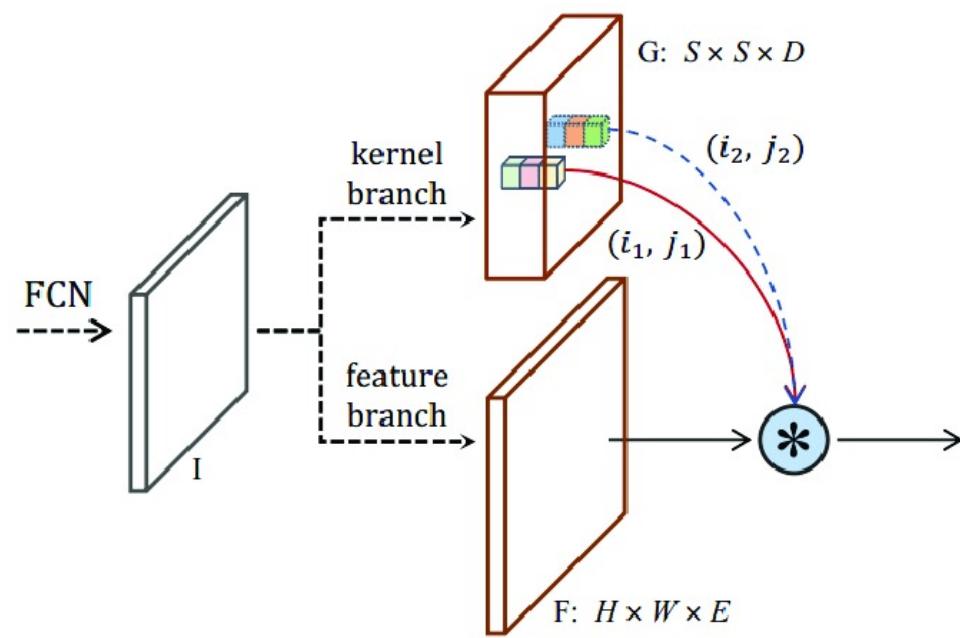


BETTER IN TIME AND PRECISION

SOL0v2

➤ Box-free one-stage approach

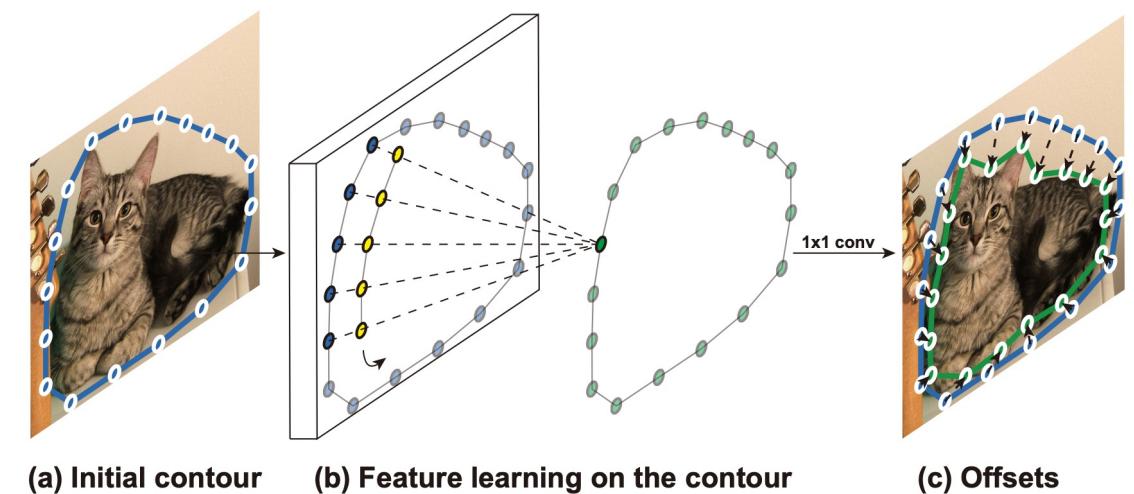
- Mask kernel prediction and mask feature learning
- Matrix non-maximum suppression (NMS)



DEEP SNAKE

➤ Contour-based approach

- Initial contour proposal iteratively deforms
- Circular Convolution

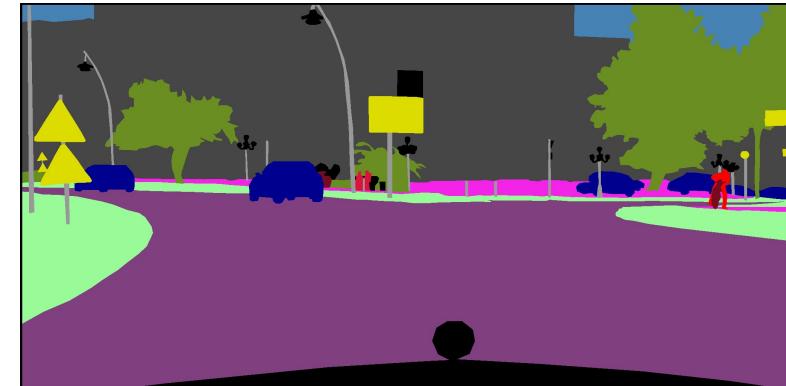


➤ Cityscapes

- Benchmarks and large-scale datasets for semantic urban scene understanding
- Video sequences, recorded in 50 different cities
- High-quality annotations or/and coarse annotation
- gtFine & leftImg8

category	#instances	category	#instances	category	#instances
person	17918	rider	1781	car	26963
truck	484	bus	380	train	168
motorcycle	737	bicycle	3675		
total	52106				

- 3475 train/val images
- 1525 test images
- 5000 total images
- 8 classes



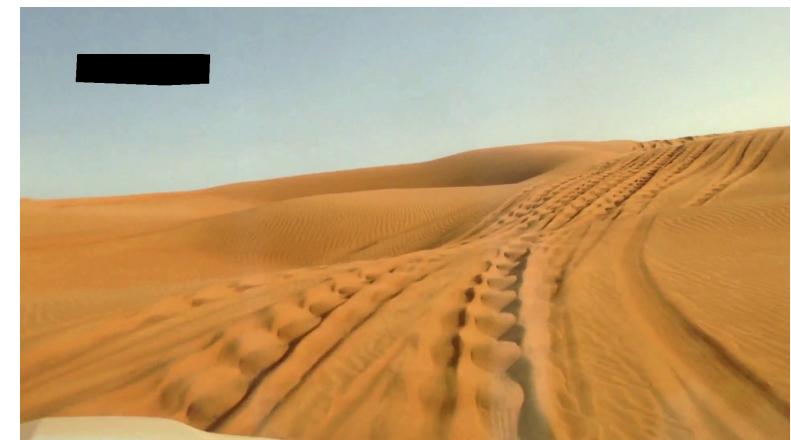
! **PROBLEM** : always images in good weather conditions

➤ WildDash

- Benchmark suite and dataset for semantic and instance segmentation for the automotive domain
- Scenarios such as rain, darkness and road cover
- public gt package

category	#instances	category	#instances	category	#instances
ego vehicle	504	person	1479	rider	589
car	4045	truck	439	bus	128
caravan	16	trailer	20	train	6
motorcycle	536	bicycle	74	pickup	266
van	201				
total	8303				

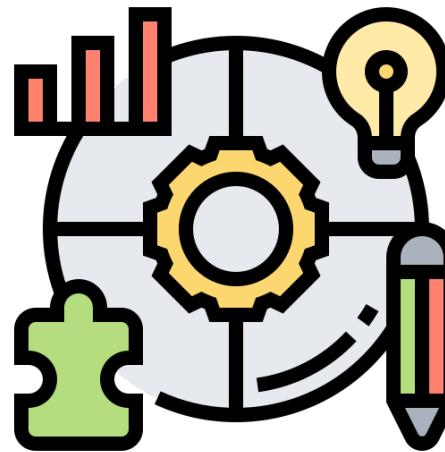
- **3405** train images
- **851** test images
- **4256** total images
- **13** classes



CHALLENGES

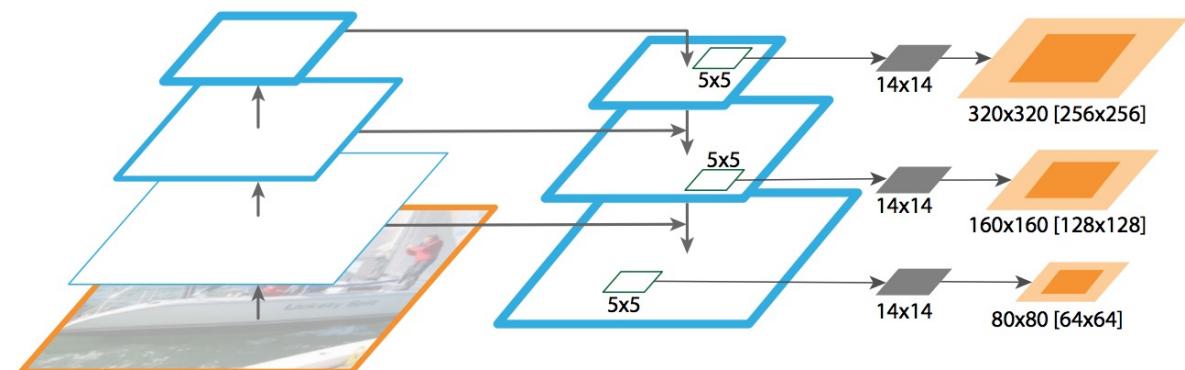
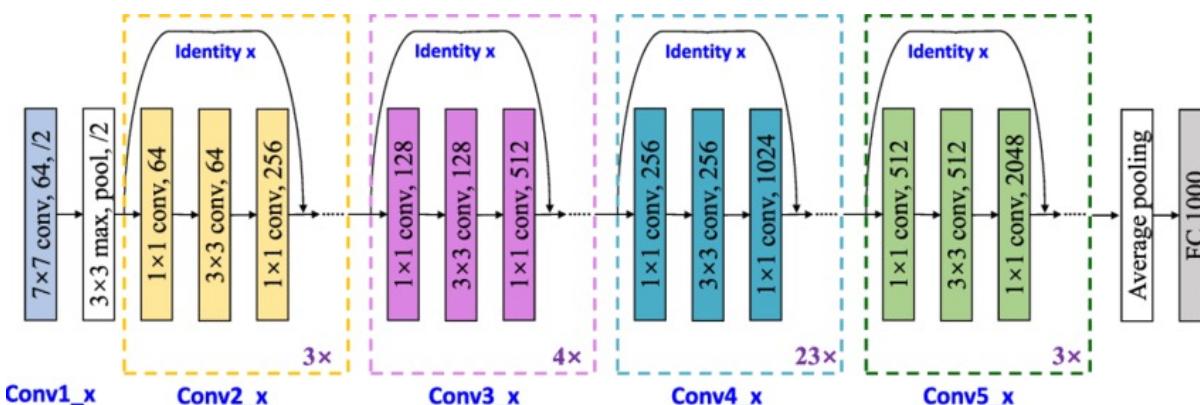
➤ Method

- Detectron2
- AdelaiDet
- Ideal models found during the experiments
- Metrics for concretise the concept of comparison

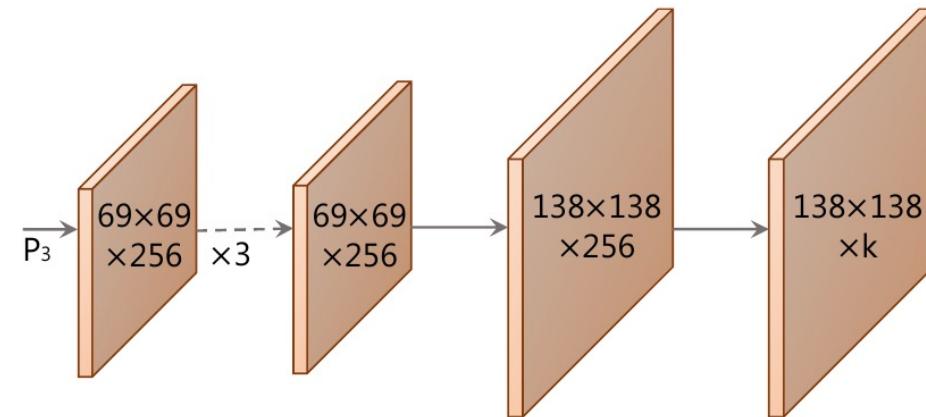


➤ Architecture

- We decided to use ResNet-101-FPN as a backbone



- In addition, as the bottom module of BlendMask, we have instantiated ProtoNet



➤ Hyperparameters

- We tried to keep the same parameters defined in the papers (Mask R-CNN & BlendMask)
- We modified those that did not allow us to run on Colab:
 - **Number of iterations:** 4.000 for Mask R-CNN (default 40.000)
 - **Number of iterations:** 2.000 for BlendMask (default 40.000)
 - **Batch size per image:** 128 (default 512)
 - **Images per batch:** 1 (we use only 1 GPU)



➤ Loss functions

Mask R-CNN

$$L = L_{cls} + L_{box} + L_{mask}$$

- L_{cls} : classification loss
- L_{box} : bounding-box loss
- L_{mask} : average binary cross-entropy loss

BlendMask

$$L^s(\alpha, p) \propto -\log \sum_{x \models \alpha} \prod_{i:x \models X_i} p_i \prod_{i:x \models \neg X_i} (1 - p_i)$$

- α : sentence in propositional logic defined on variables X_1, \dots, X_n
- p : probability vector for each variable X_i
- $L^s(\alpha, p)$: semantic loss between α and p

➤ Evaluation

- We use Average Precision (AP):

$$\int_0^1 p(r)dr$$

- $p(r)$: the area under the curve of maximum intersection between:

- Recall

$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

- Precision

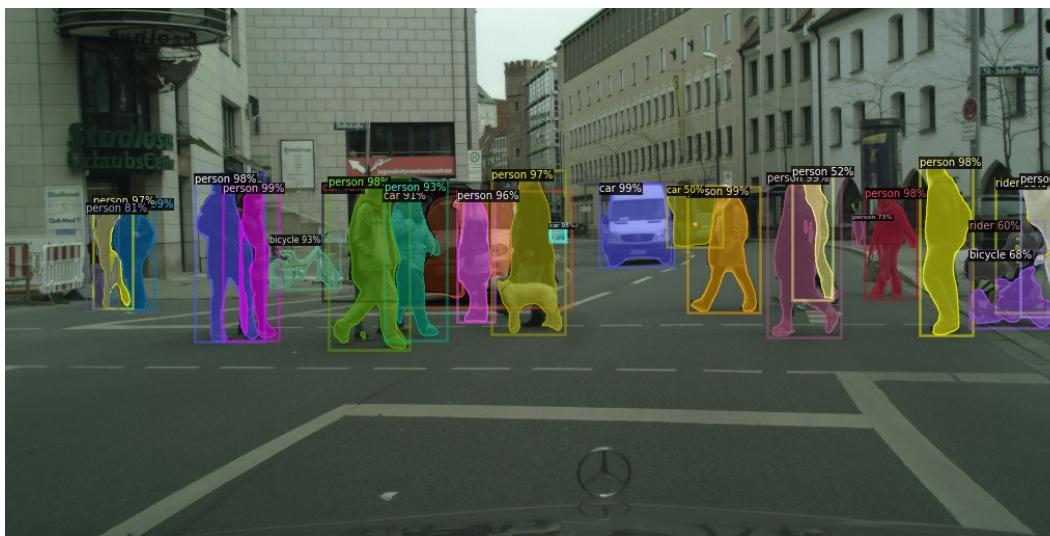
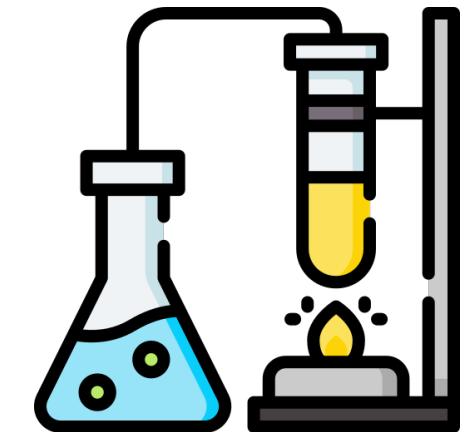
$$\frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

- AP is averaged over Intersection over Union (IoU) values
- 10 IoU thresholds from the range .50:.05:.95
- from these thresholds are extracted the 100 detections with the highest score



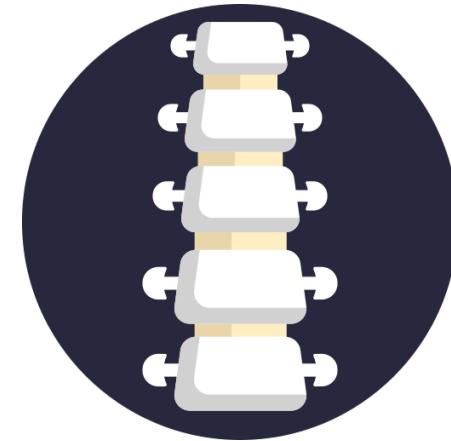
➤ Experiments

- Define "Ideal Models" via experiments:
 - Backbones
 - Frozen Layer
 - (Parameters to run on Colab)



➤ Backbones

Architecture	<i>Cityscapes</i>	<i>WildDash</i>
Mask R-CNN: ResNet-50-C4	box AP: 25.935 mask AP: 20.005	box AP: 18.730 mask AP: 17.589
Mask R-CNN: ResNet-50-DC5	box AP: 25.591 mask AP: 20.285	box AP: 17.871 mask AP: 16.543
Mask R-CNN: ResNet-50-FPN	box AP: 24.437 mask AP: 20.327	box AP: 16.865 mask AP: 16.017
Mask R-CNN: ResNet-101-C4	box AP: 30.888 mask AP: 24.562	box AP: 20.459 mask AP: 19.331
Mask R-CNN: ResNet-101-DC5	box AP: 29.125 mask AP: 24.062	box AP: 19.391 mask AP: 18.529
Mask R-CNN: ResNet-101-FPN	box AP: 28.563 mask AP: 24.676	box AP: 19.024 mask AP: 17.994

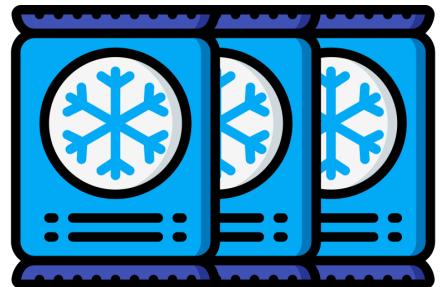
Table 1. Experiments on *backbone* Mask R-CNN.

Architecture	<i>Cityscapes</i>	<i>WildDash</i>
BlendMask: ResNet-50-FPN	box AP: 30.129 mask AP: 25.325	box AP: 20.225 mask AP: 19.810
BlendMask: ResNet-101-FPN	box AP: 31.017 mask AP: 26.411	box AP: 20.347 mask AP: 20.162
BlendMask: ResNet-101-FPN + dcni	box AP: 32.196 mask AP: 27.212	box AP: 20.497 mask AP: 20.236

Table 2. Experiments on *backbone* BlendMask.

ResNet-101-FPN

➤ Frozen Layers



Architecture	<i>Cityscapes</i>	<i>WildDash</i>
Mask R-CNN: ResNet-101-FPN + freeze at 2nd	box AP: 28.563 mask AP: 24.676	box AP: 19.024 mask AP: 17.994
Mask R-CNN: ResNet-101-FPN + freeze at 3rd	box AP: 26.297 mask AP: 22.179	box AP: 19.132 mask AP: 18.224
Mask R-CNN: ResNet-101-FPN + freeze at 4th	box AP: 25.053 mask AP: 21.349	box AP: 19.824 mask AP: 18.448

Table 3. Experiments on *frozen layers* Mask R-CNN.

Architecture	<i>Cityscapes</i>	<i>WildDash</i>
BlendMask: ResNet-101-FPN + freeze at 2nd	box AP: 31.017 mask AP: 26.411	box AP: 20.347 mask AP: 20.162
BlendMask: ResNet-101-FPN + freeze at 3rd	box AP: 28.596 mask AP: 23.649	box AP: 18.993 mask AP: 17.802
BlendMask: ResNet-101-FPN + freeze at 4th	box AP: 31.379 mask AP: 26.936	box AP: 21.597 mask AP: 20.537

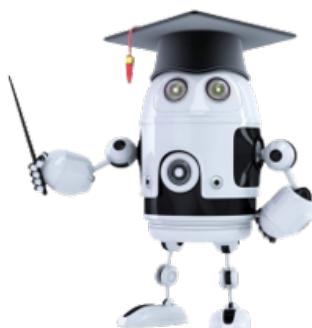
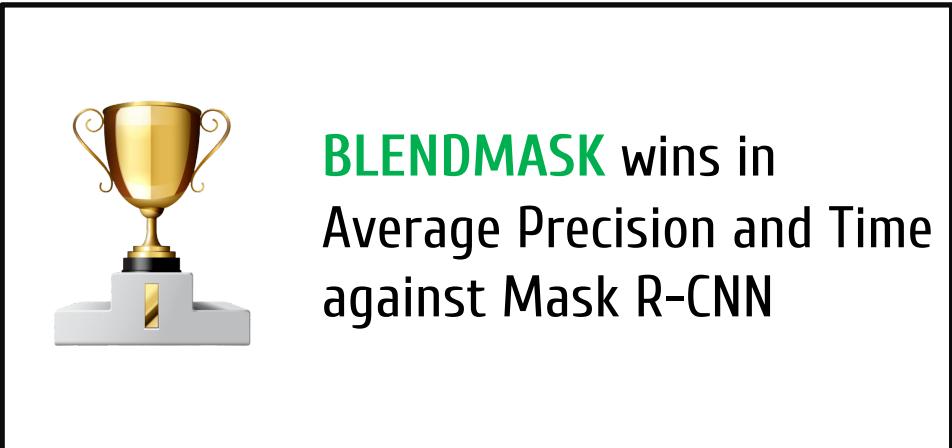
Table 4. Experiments on *frozen layers* BlendMask.

➤ Ideal Models

Method	<i>Cityscapes</i>	<i>WildDash</i>
Mask R-CNN:	box AP: 28.563 mask AP: 24.676	box AP: 19.024 mask AP: 17.994
BlendMask:	box AP: 31.017 (+2.45 AP) mask AP: 26.411 (+1.74 AP)	box AP: 20.347 (+1.32 AP) mask AP: 20.162 (+2.17 AP)

➤ Conclusion

- One-stage and anchor box-free techniques, suitably modified, perform better than box-based two-stage methods
- Caused by the hybridisation of top-down and bottom-up methods and the detection of unanchored objects

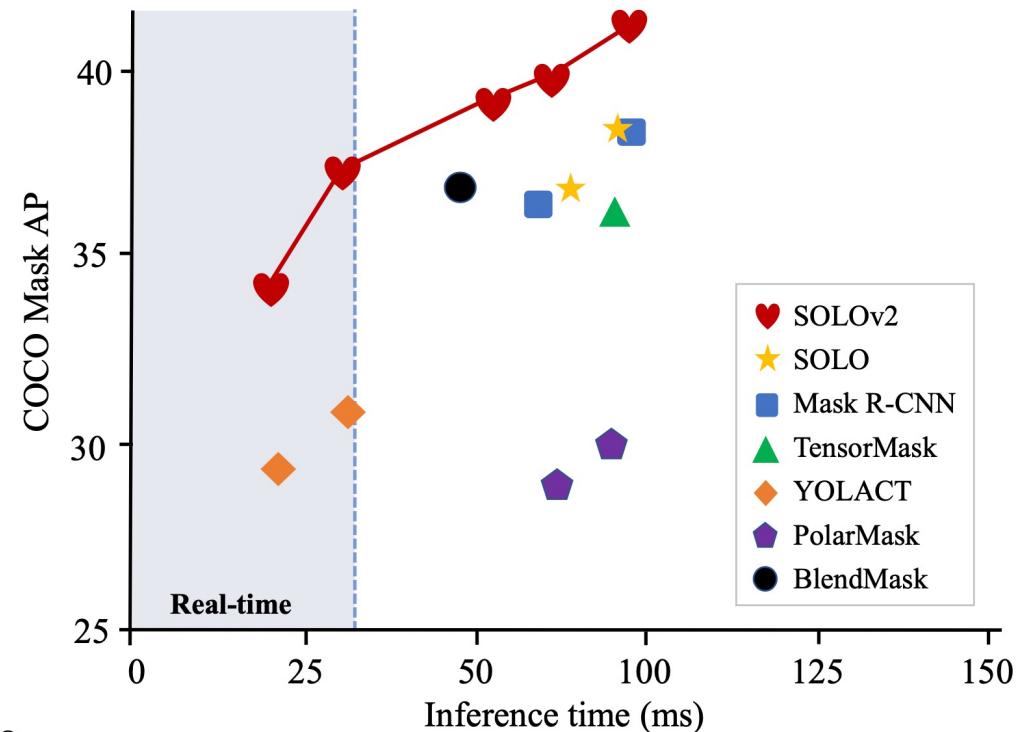


Is it possible to do better ?

- A box-free stage approach combined with matrix NMS, such as **SOL0v2**, is demonstrated to be very competitive against BlendMask

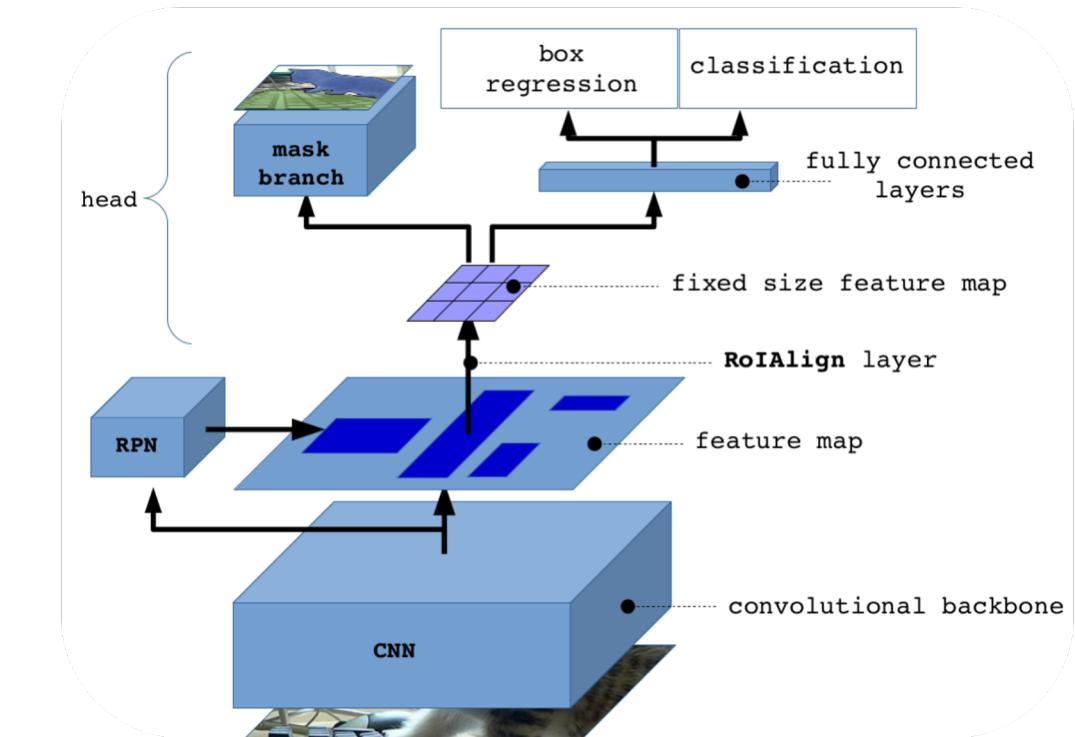
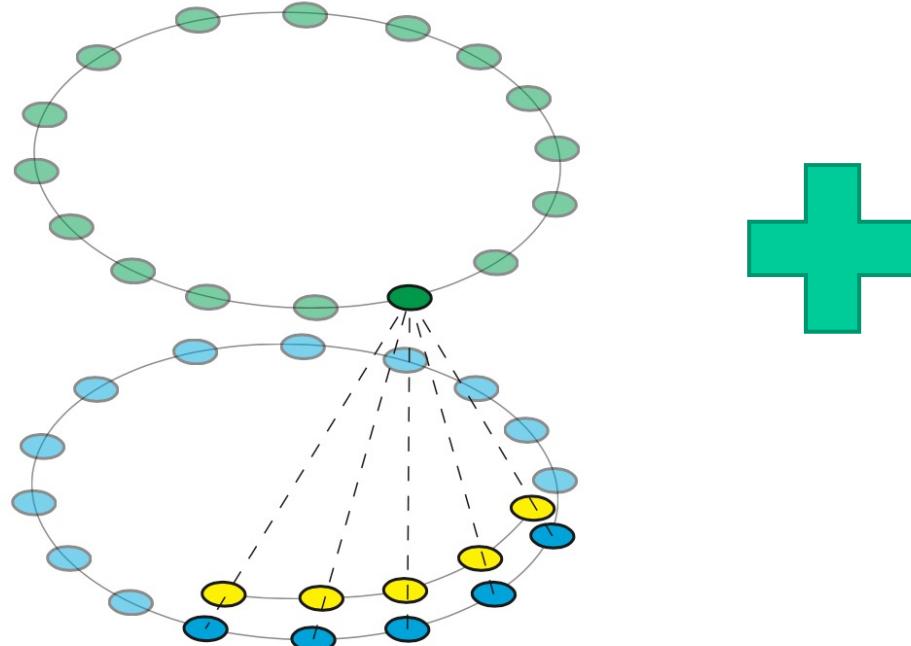
- Deep Snake outperforms Mask R-CNN in inference speed and AP

	training data	fps	AP [val]	AP	AP ₅₀
SGN [26]	fine + coarse	0.6	29.2	25.0	44.9
PolygonRNN++ [1]	fine	-	-	25.5	45.5
Mask R-CNN [18]	fine	2.2	31.5	26.2	49.9
GMIS [28]	fine + coarse	-	-	27.6	49.6
Spatial [31]	fine	11	-	27.6	50.9
PANet [27]	fine	<1	36.5	31.8	57.1
Deep snake	fine	4.6	37.4	31.7	58.4



➤ Possible extensions

- Improve the performance of BlendMask by trying to use Deep Snake contour-based approach for detecting object bounding boxes of an image





That's all folks!

Francesco Bari

francesco.bari.2@studenti.unipd.it

Eleonora Signor

eleonora.signor@studenti.unipd.it