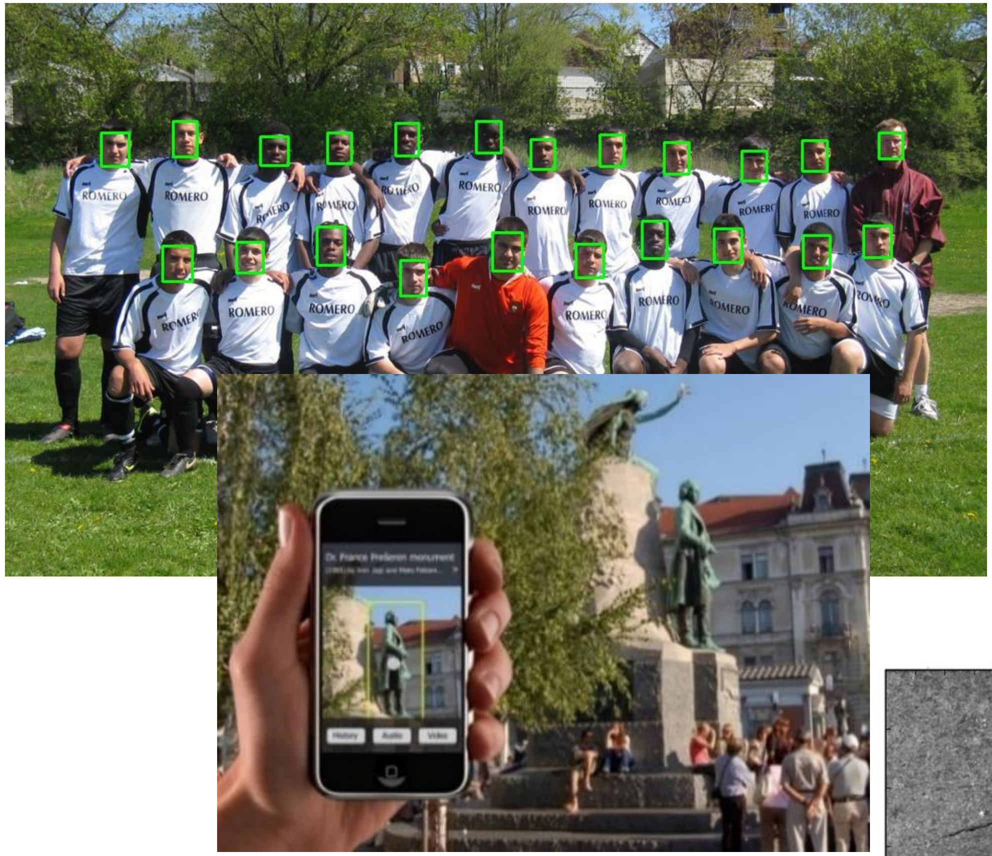
Deep Learning

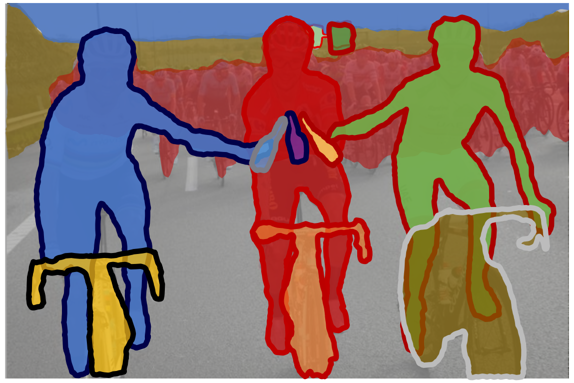
Computer vision beyond classification

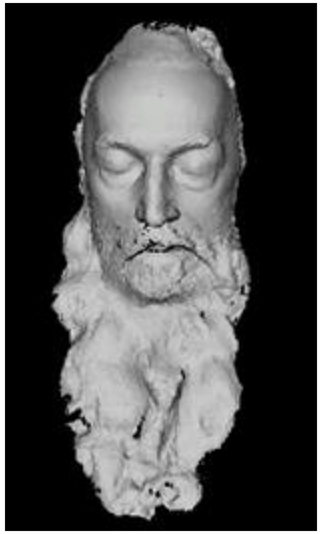
Danijel Skočaj
University of Ljubljana
Faculty of Computer and Information Science

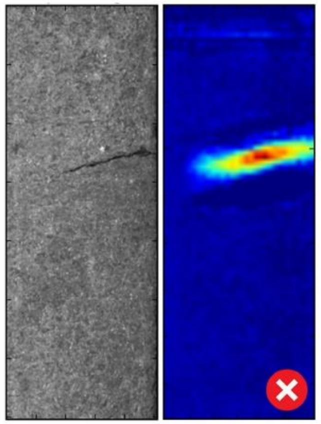
Academic year: 2022/23

Computer vision





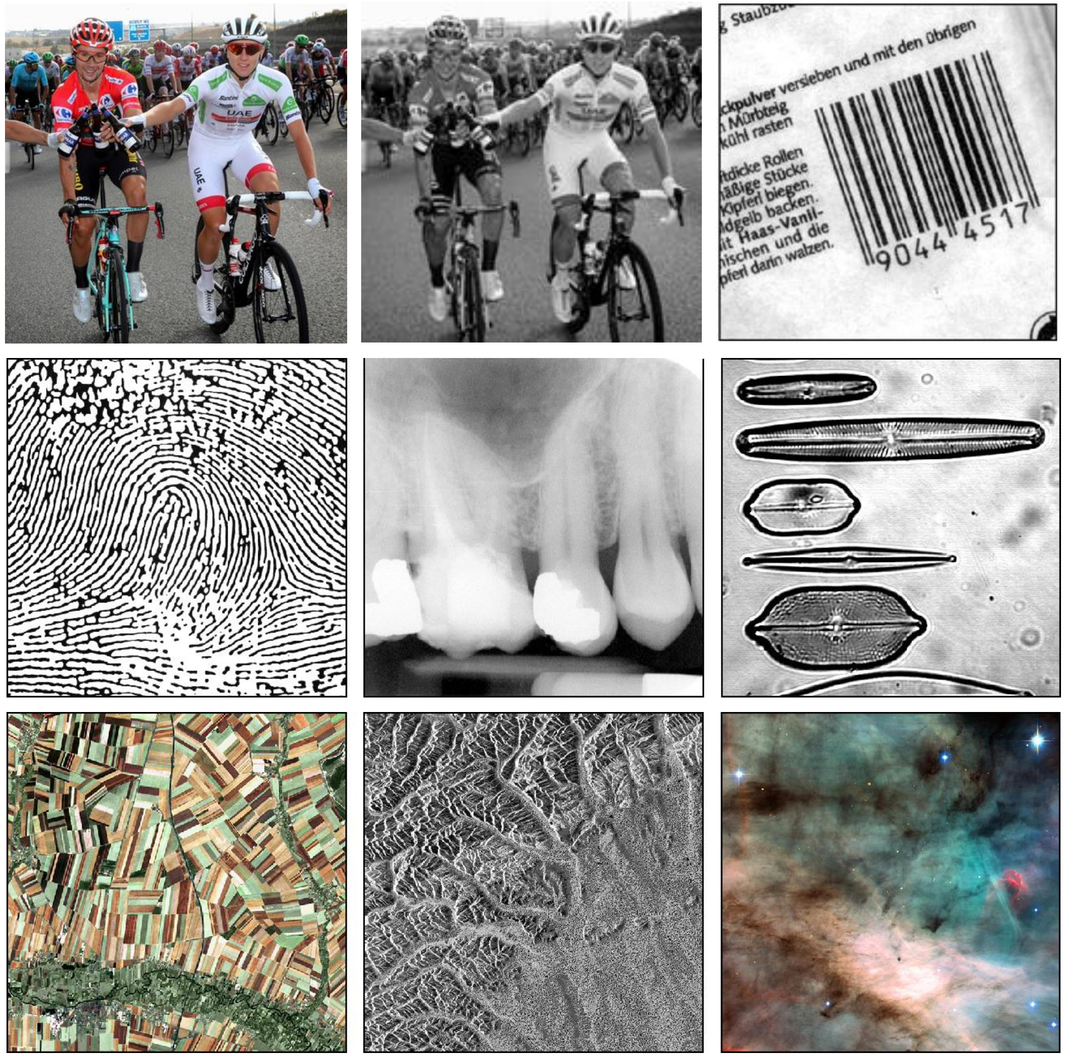




Visual information

Computer vision tasks

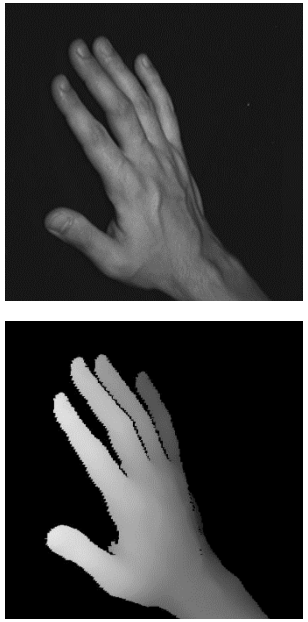
Visual information

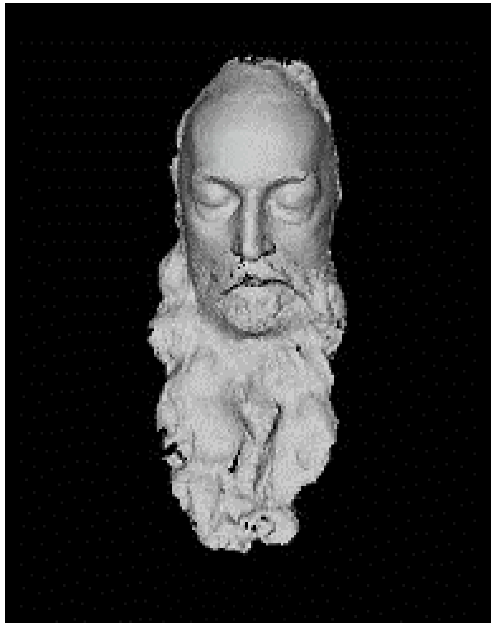


Images

Video

3D





Classification

- What is depicted in the image?

Categorisation

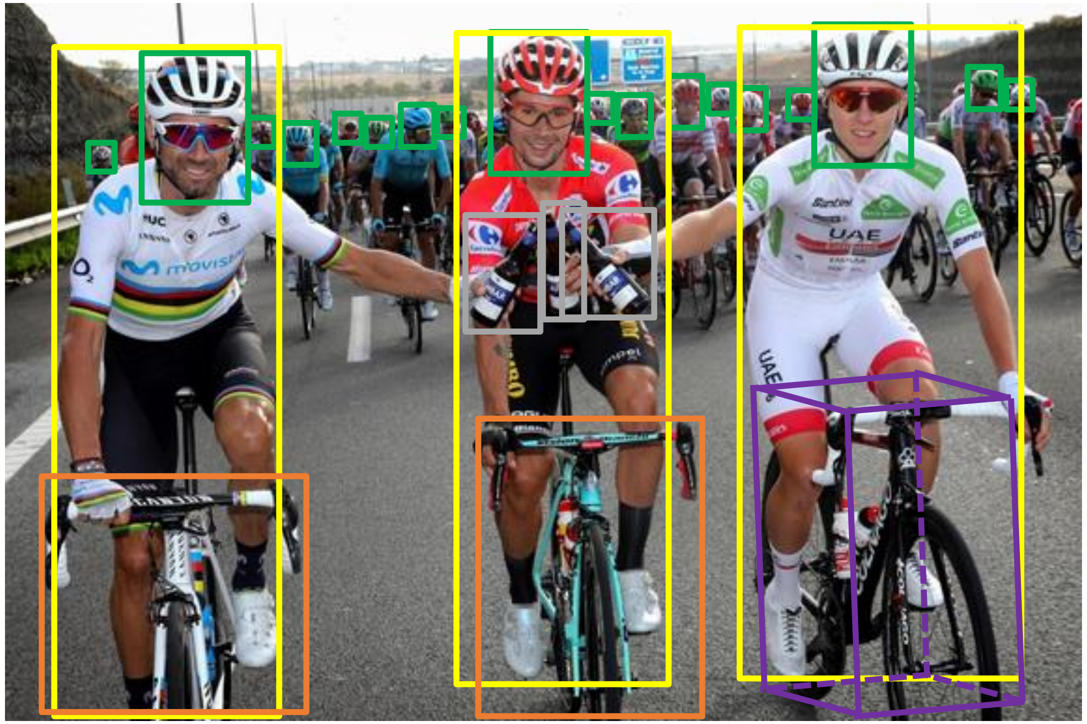
Recognition/identification of instances

Localisation

Detection

- Where in the image?

Detection



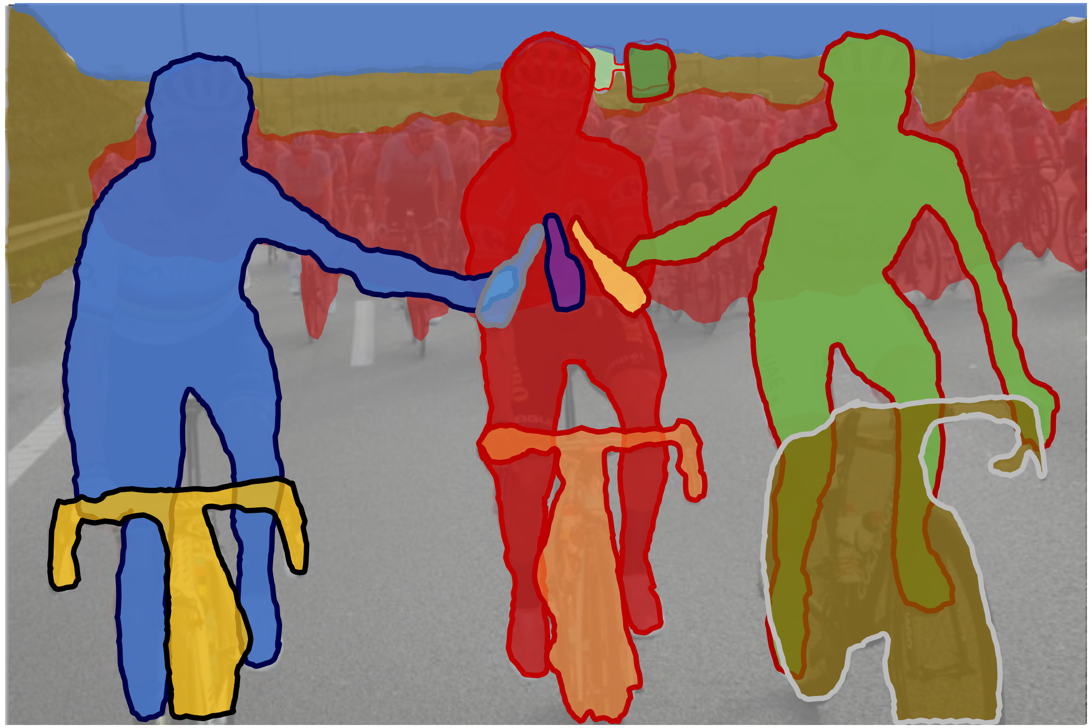
Instance segmentation

Segmentation

- What does every pixel represent?

Semantic segmentation

Panoptic segmentation



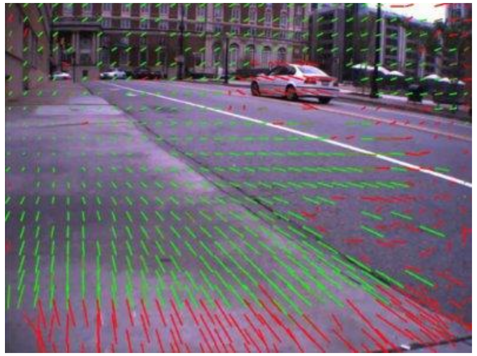
Recognition

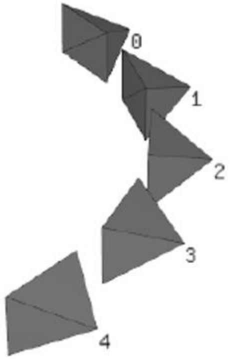
- Recognition of
  - objects
  - properties
  - faces
  - rooms
  - affordances
  - actions
  - relations
  - intentions,…
- Categorisation
- Multimodal recognition

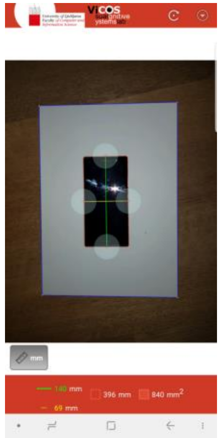
Other computer vision tasks

- Visual retrieval
- Visual tracking
- Motion analysis
- 3D computer vision
  - 3D reconstruction
  - Measurement
  - Pose estimation
- …

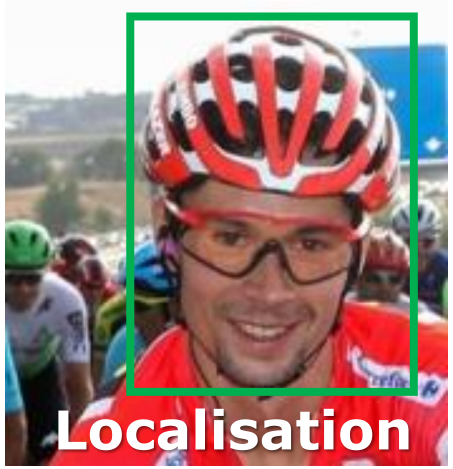


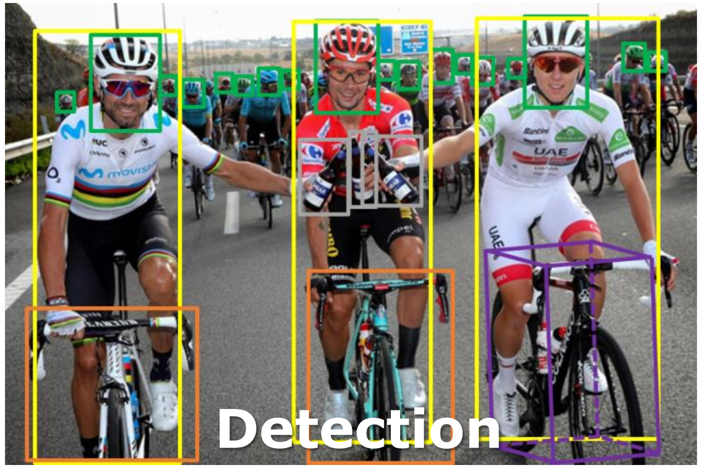



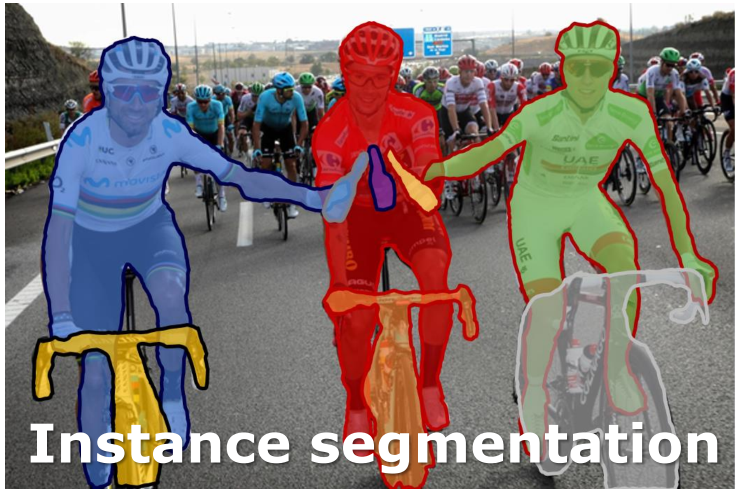



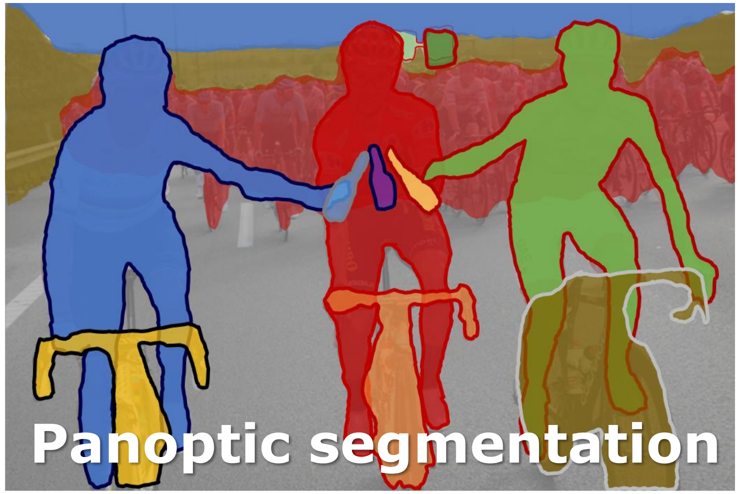


Main computer vision tasks

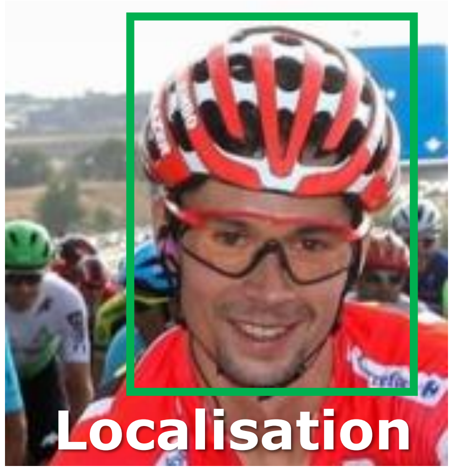


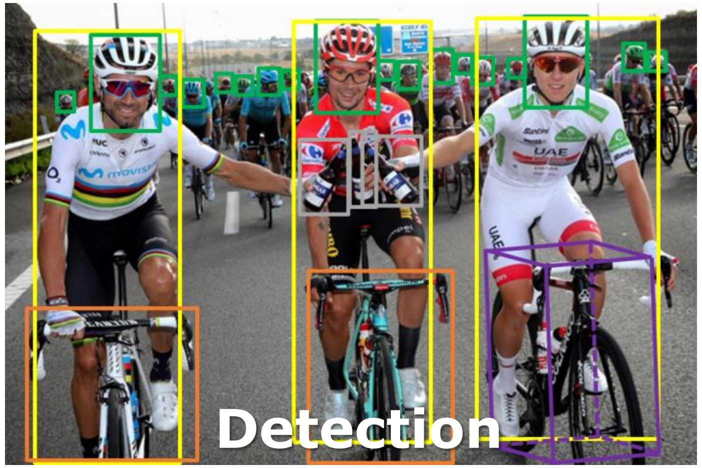


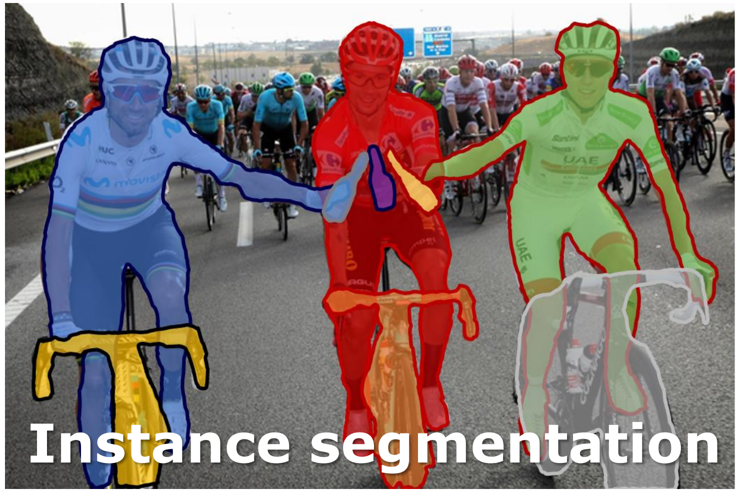


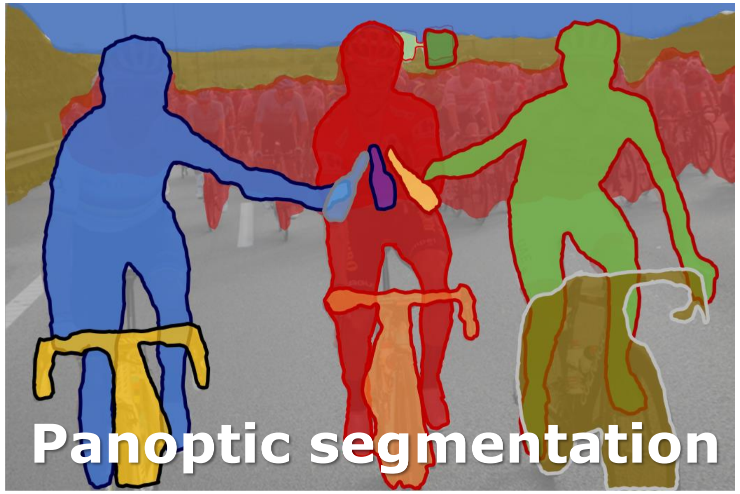


Classification



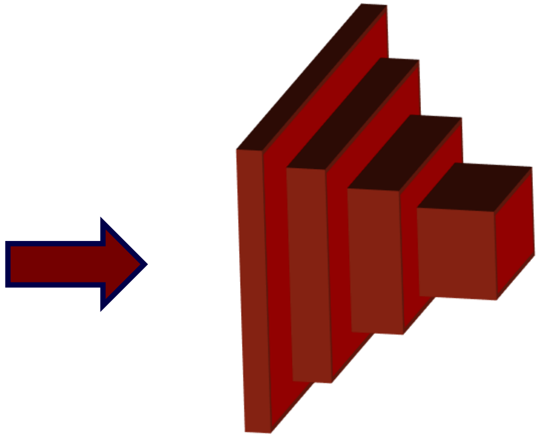




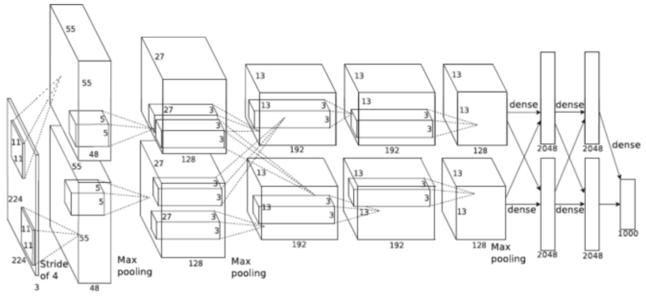


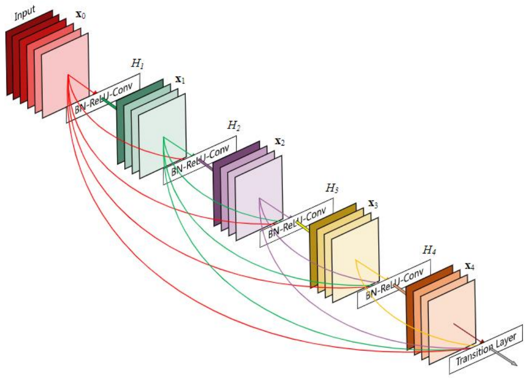
Classification

- Image classification: What is in the image?

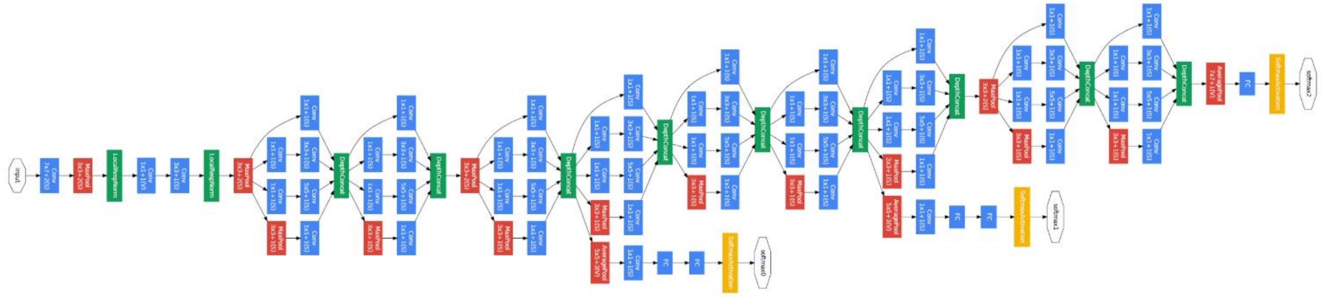


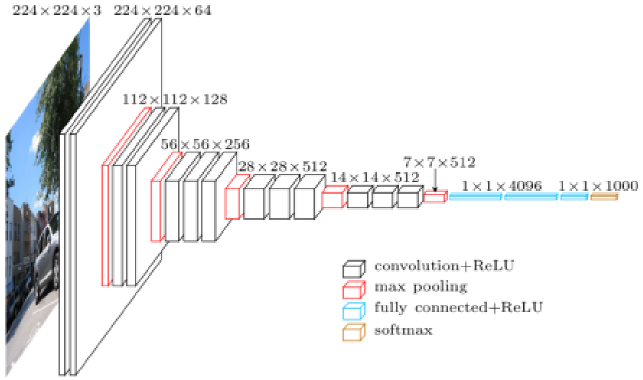
- T. Pogačar
- W. van Aert
- P. Roglič
- L. Dončić
- J. Oblak
- E. Klinec

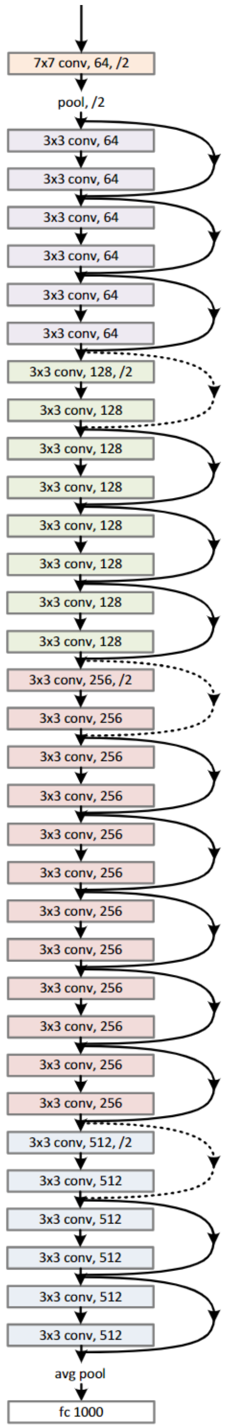




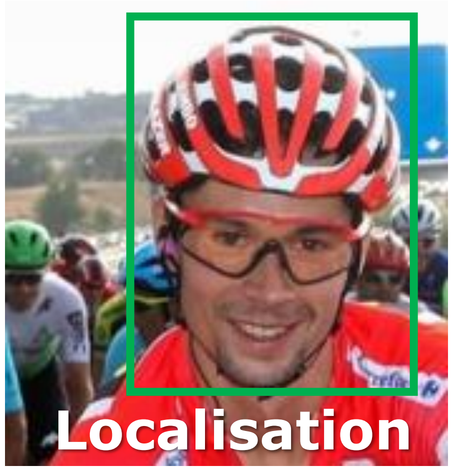
- Typically Cross entropy loss is used
- Any CNN backbone architecture can be used

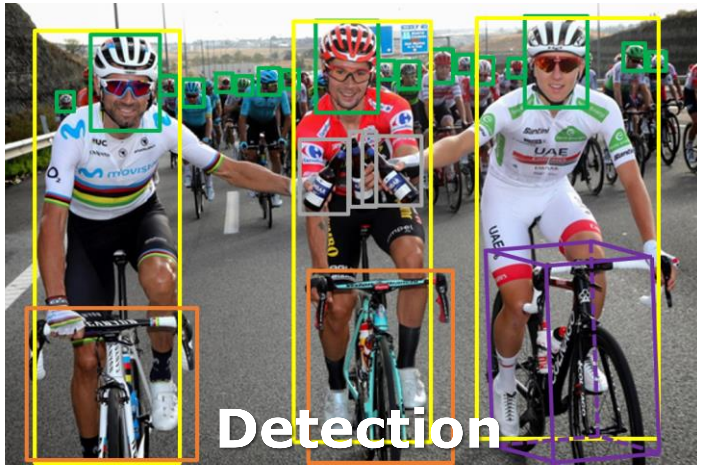


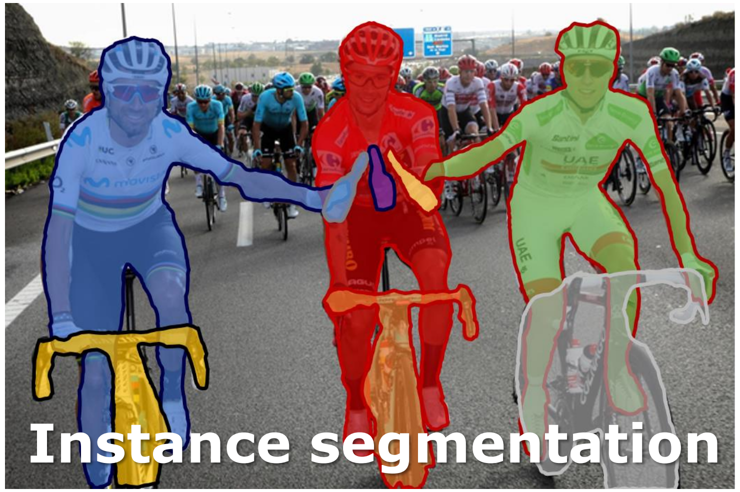


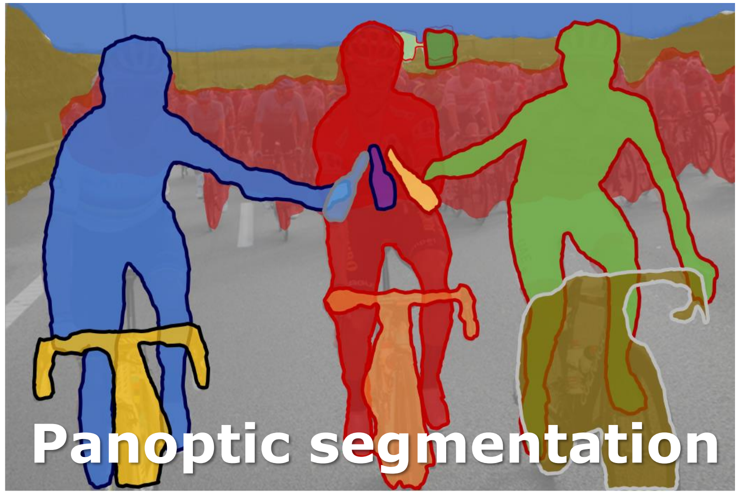


Localisation



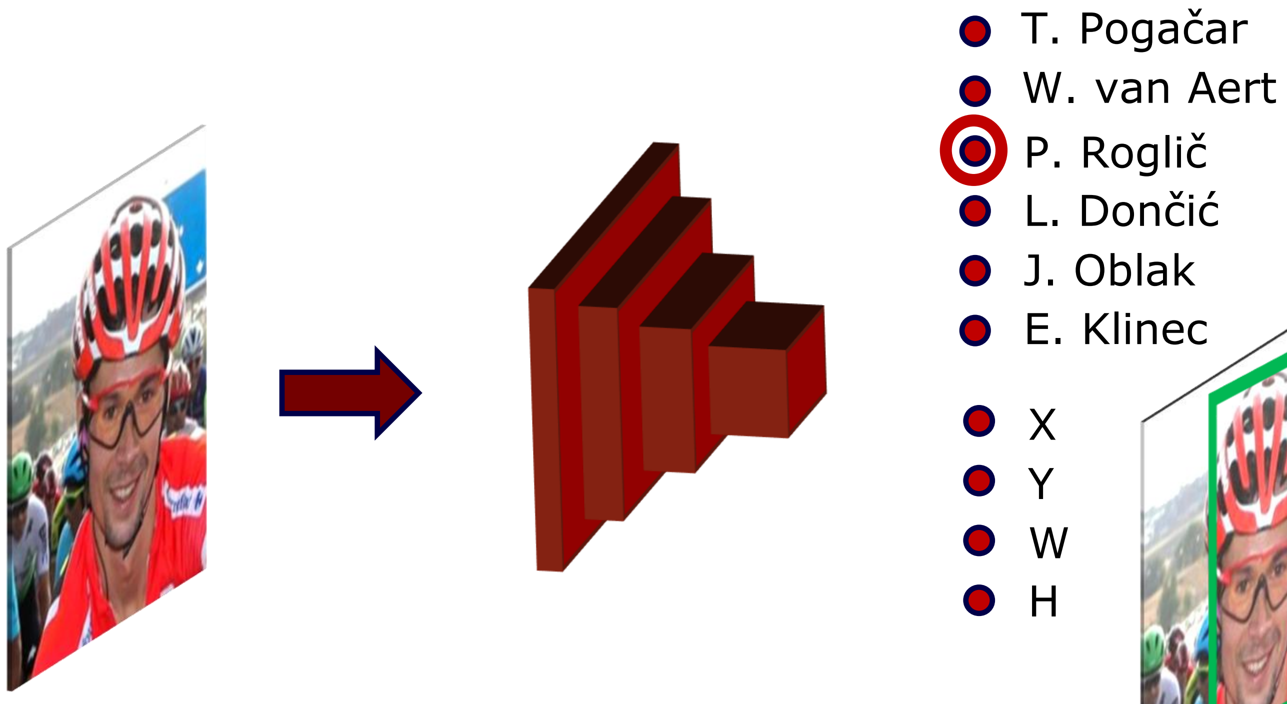


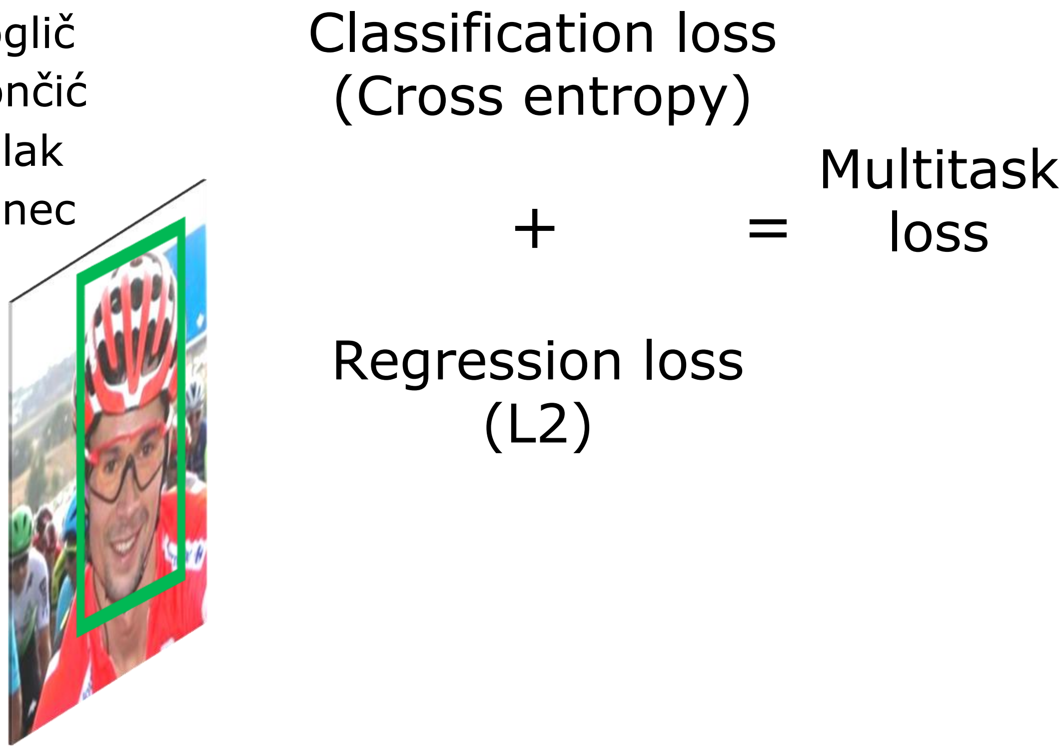




Localisation

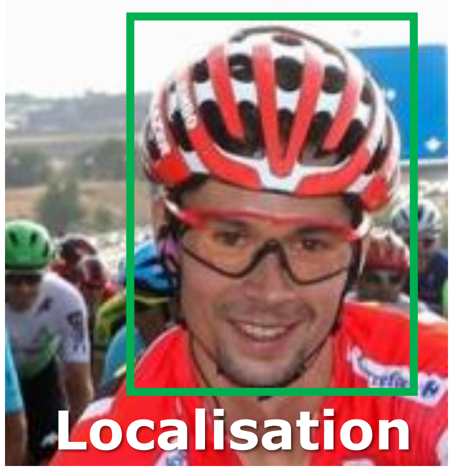
- Object localisation – Where (besides what) in the image (is the only object)?

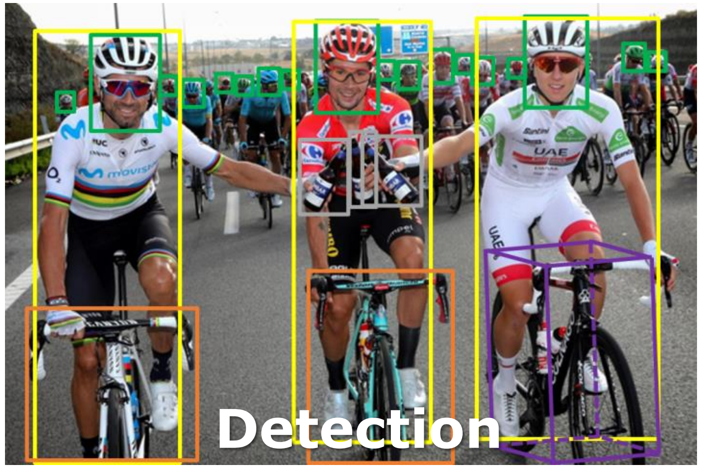


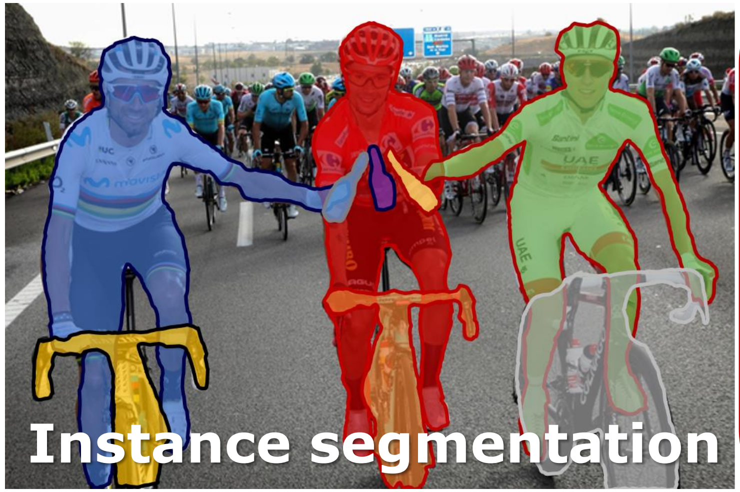


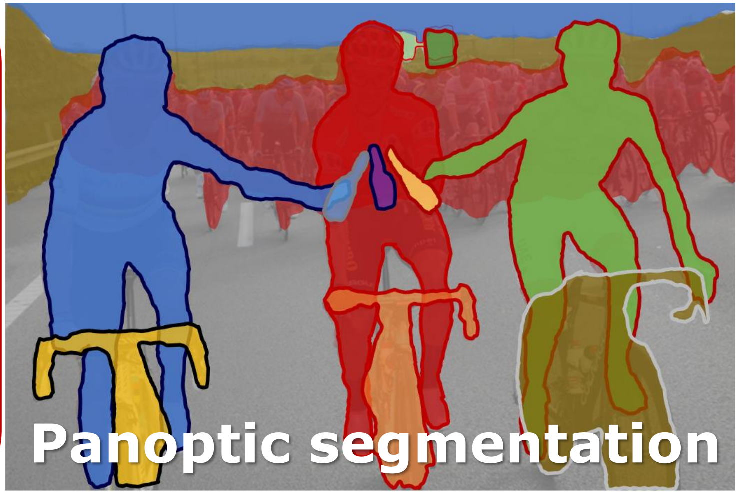
- Regress the bounding box

Semantic segmentation









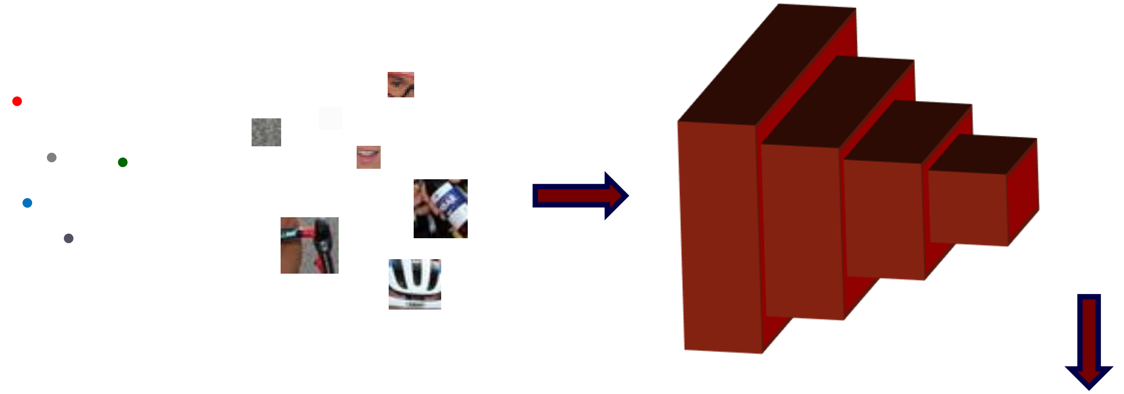
Semantic segmentation

- Classify every pixel

- Training using (image, segmentation mask) pairs

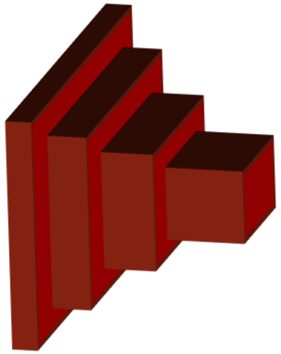
Naive approach

- Classification of every pixel
- Classification of every patch
  - Sliding window approach
- Very inefficient!

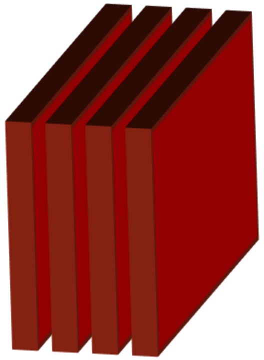


Fully convolutional approach

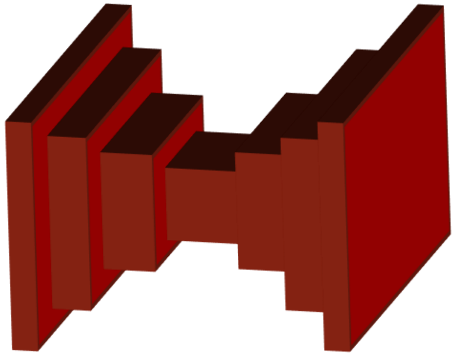
- Encoder approach
  - Downsampling
  - Small output resolution ☹



- Convolutions withouth downsampling
  - Inefficient ☹

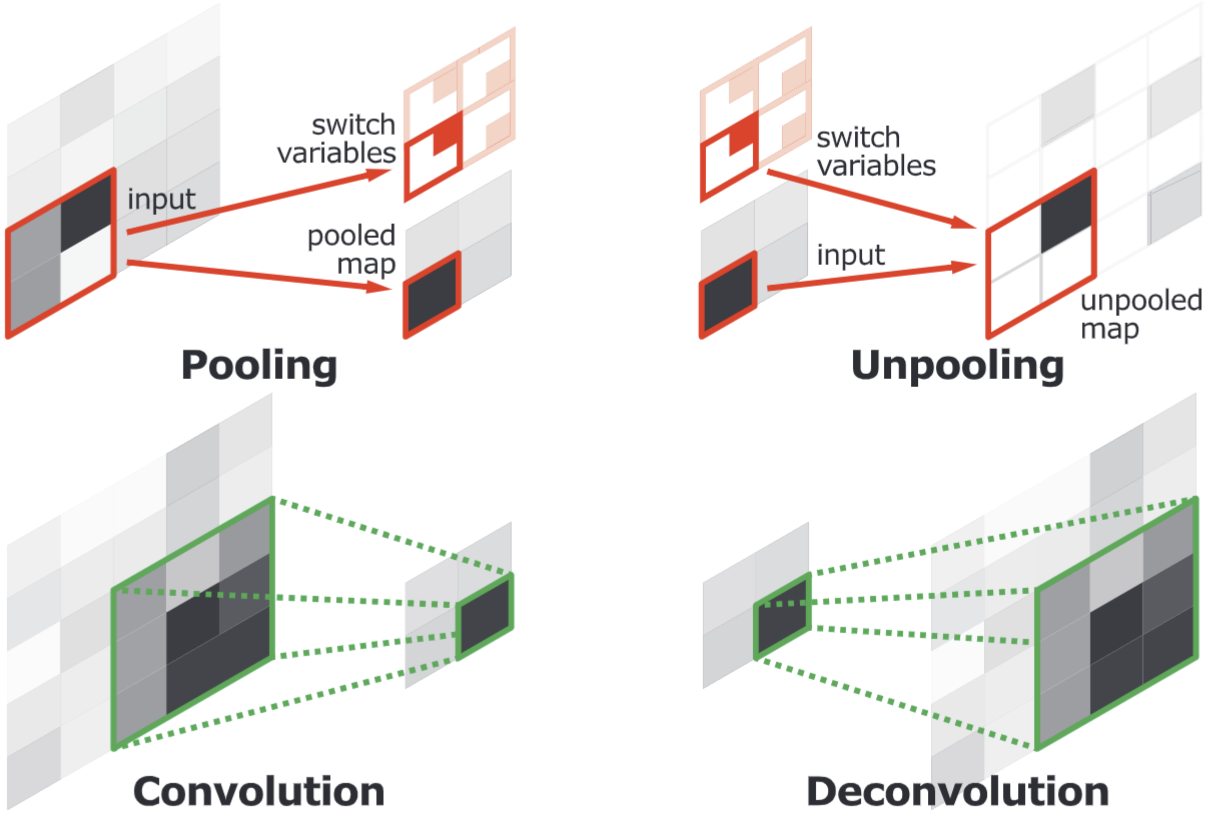


- Encoder-decoder approach
  - Downsampling + upsampling
  - High resolution ☺
  - Efficient ☺



Upsampling

- Increasing the resolution
- Nonlearnable
  - Nearest neigbour
  - Bilinear interpolation
- Unpooling
- Transpose convolution

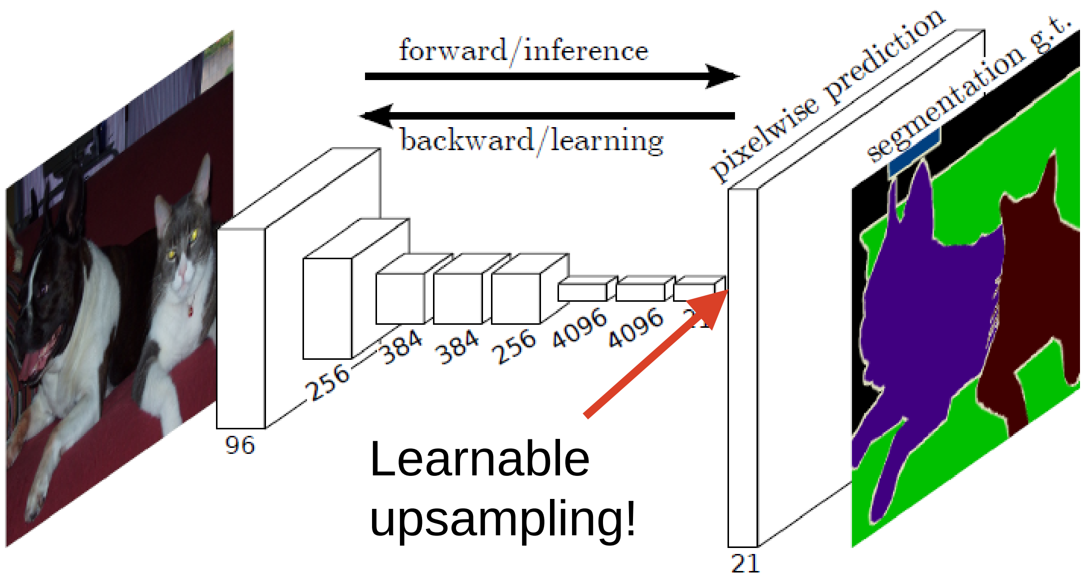


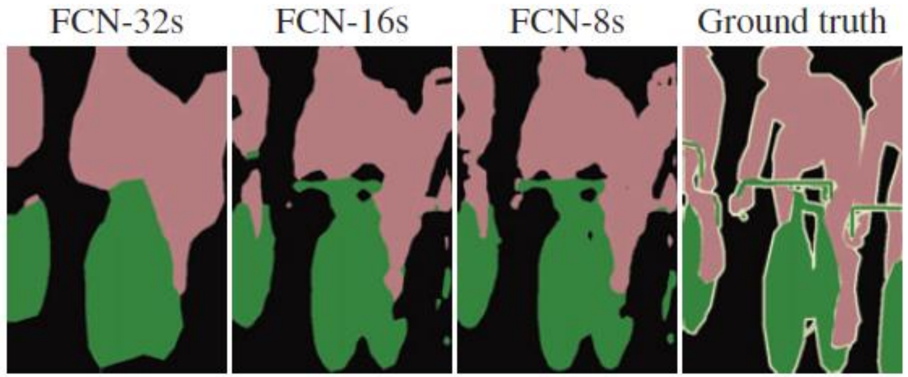
[Long et al., 2014]

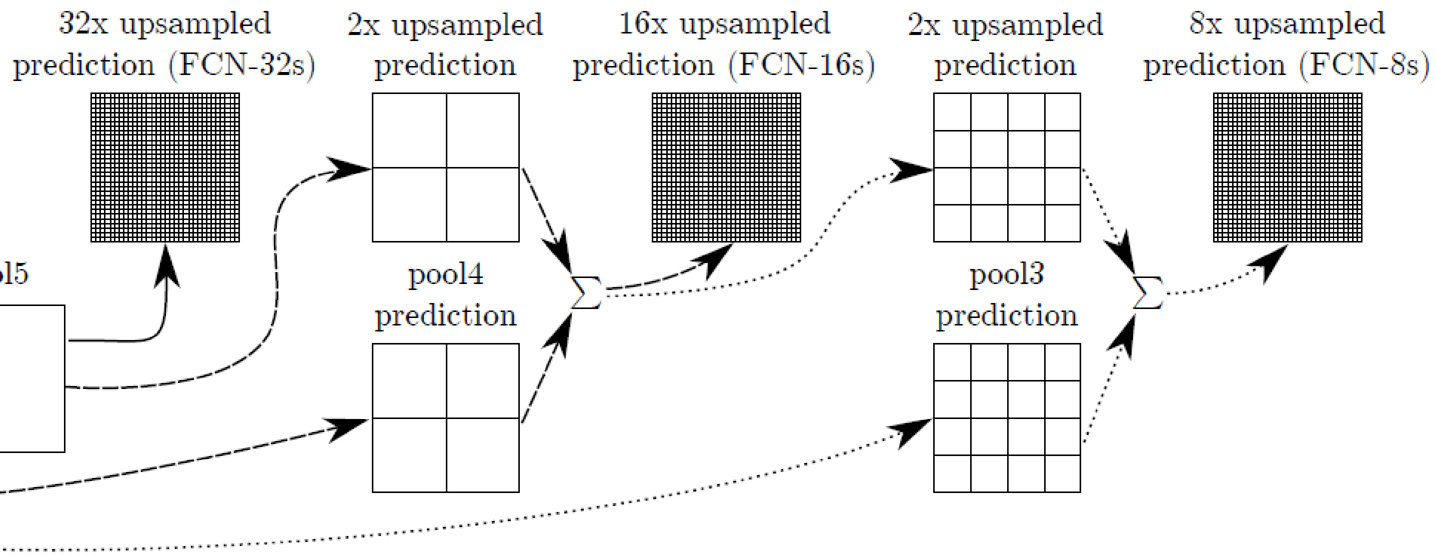
[Noh et al., 2015]

FCN

- Fully Convolutional Networks for Semantic Segmentation
- Leaernable upsampling
- Skip connections for more accurate results

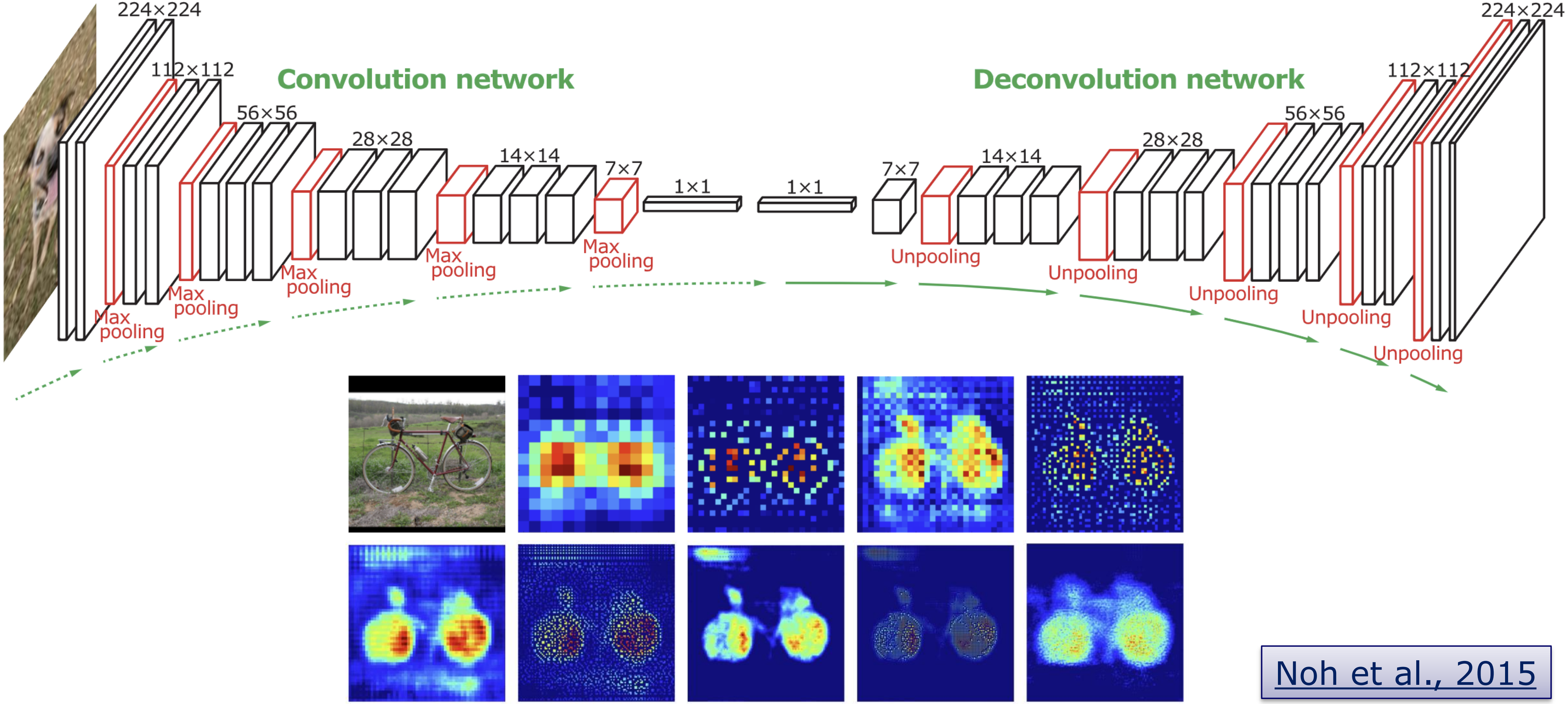






[Long et al., 2014]

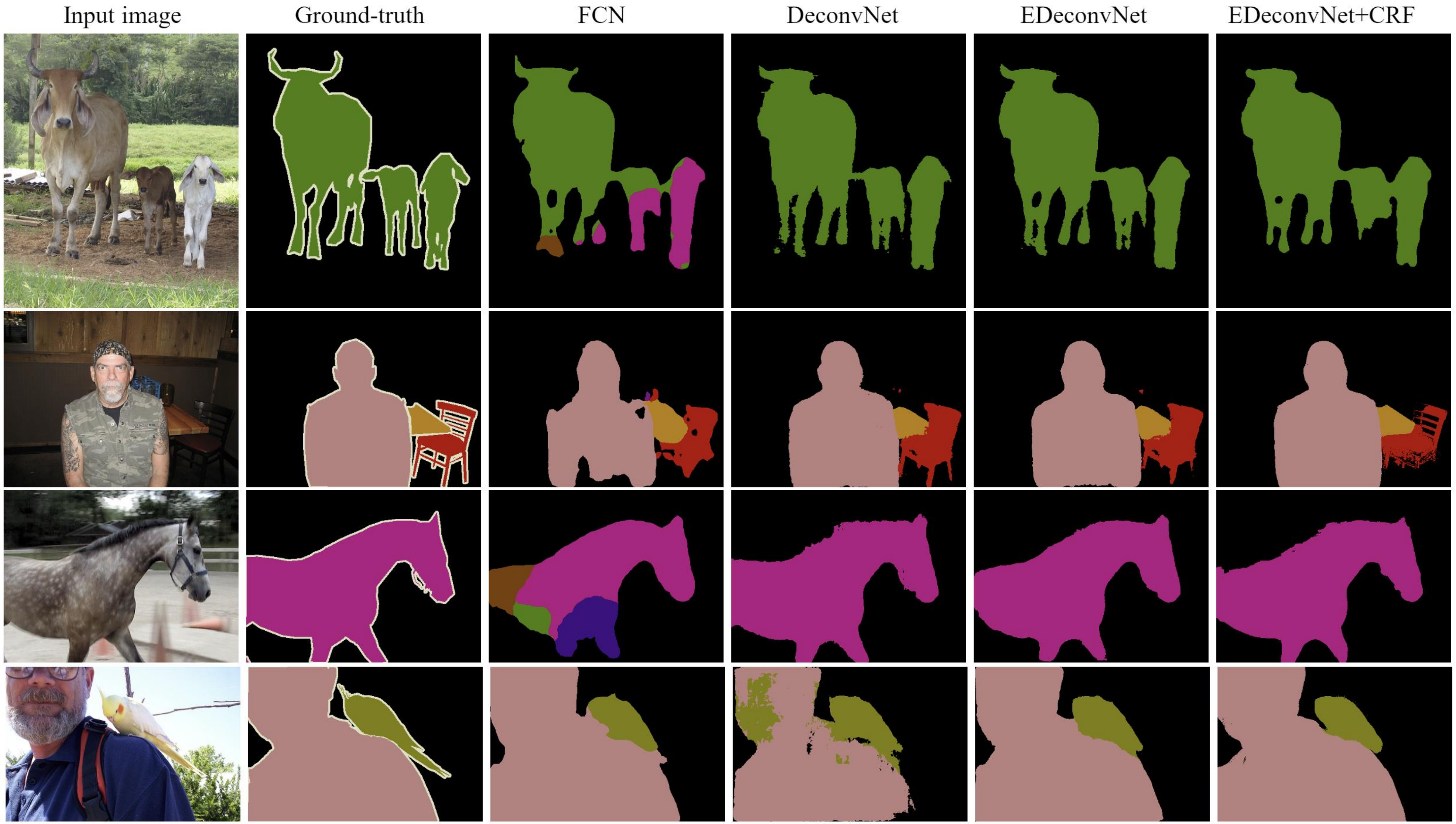
Deconvolution network



Noh et al., 2015

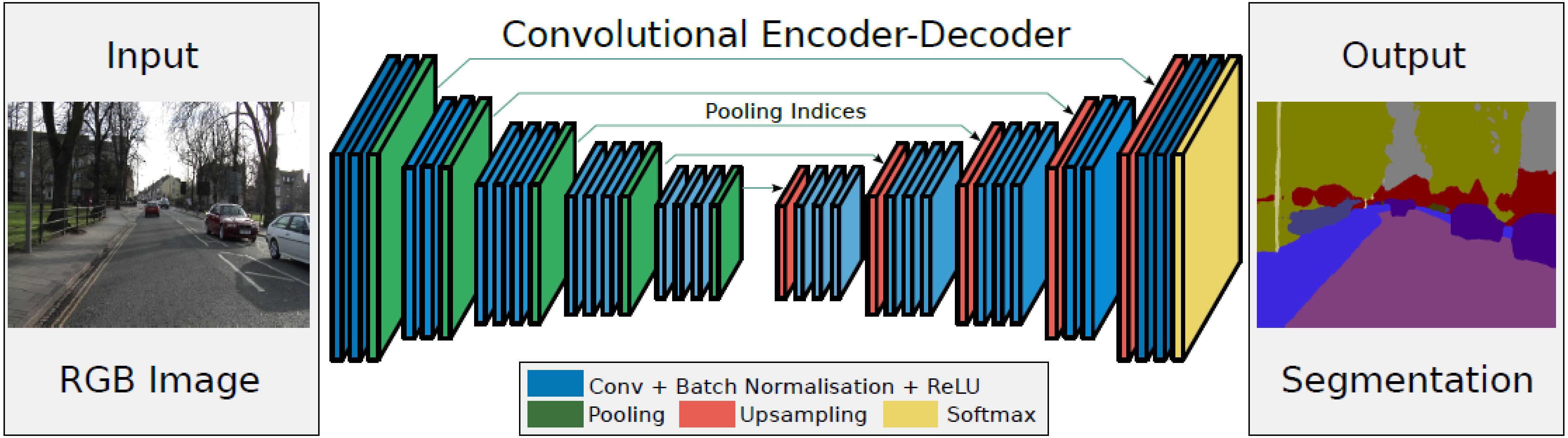
Segmentation results

Noh et al., 2015



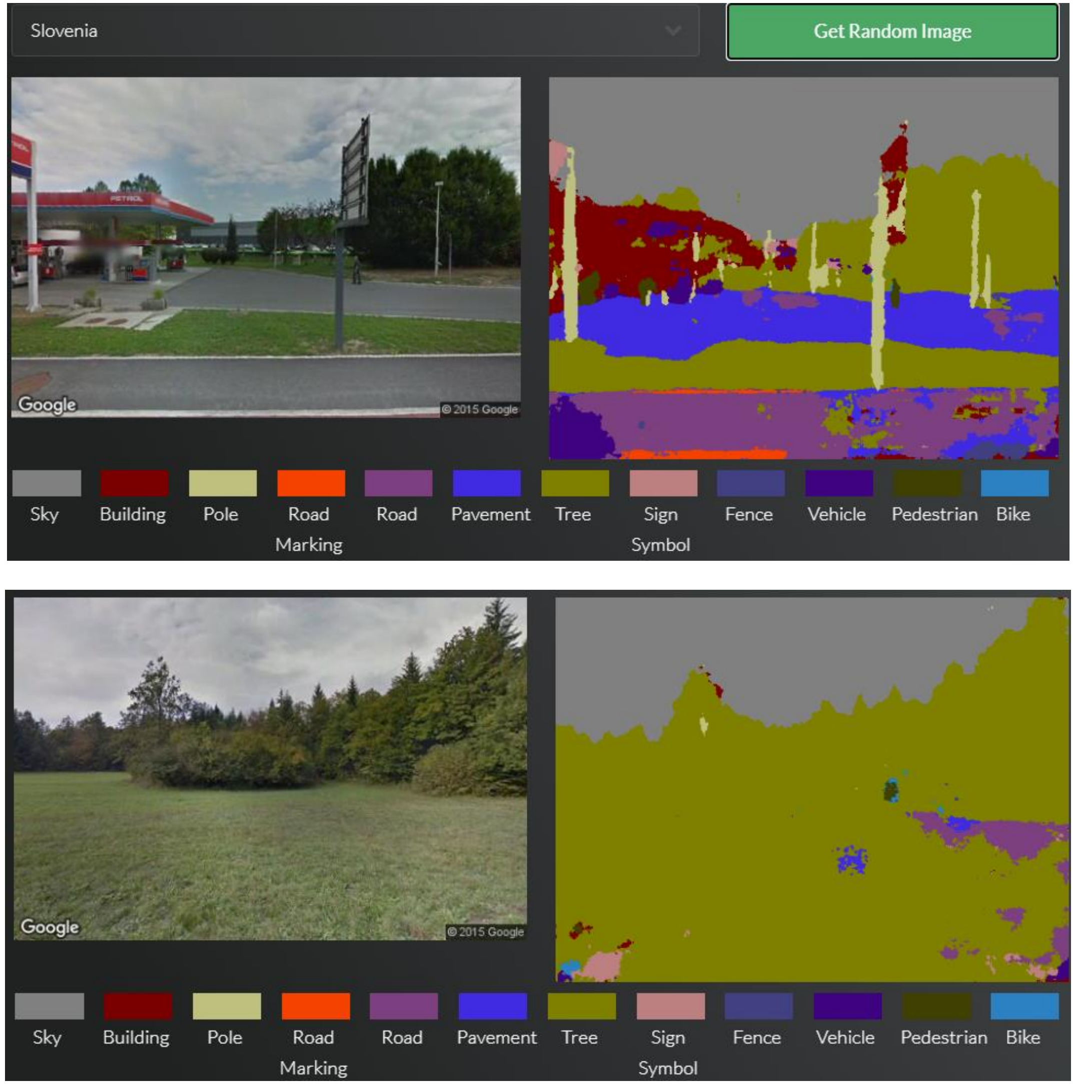
SegNet

- A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation
- Encoder (VGG16) - decoder architecture
- Upsampling with max-unpooling by storing pooling indices
- Convolutions with trainable filtes to densify activation maps
- SoftMax at the end

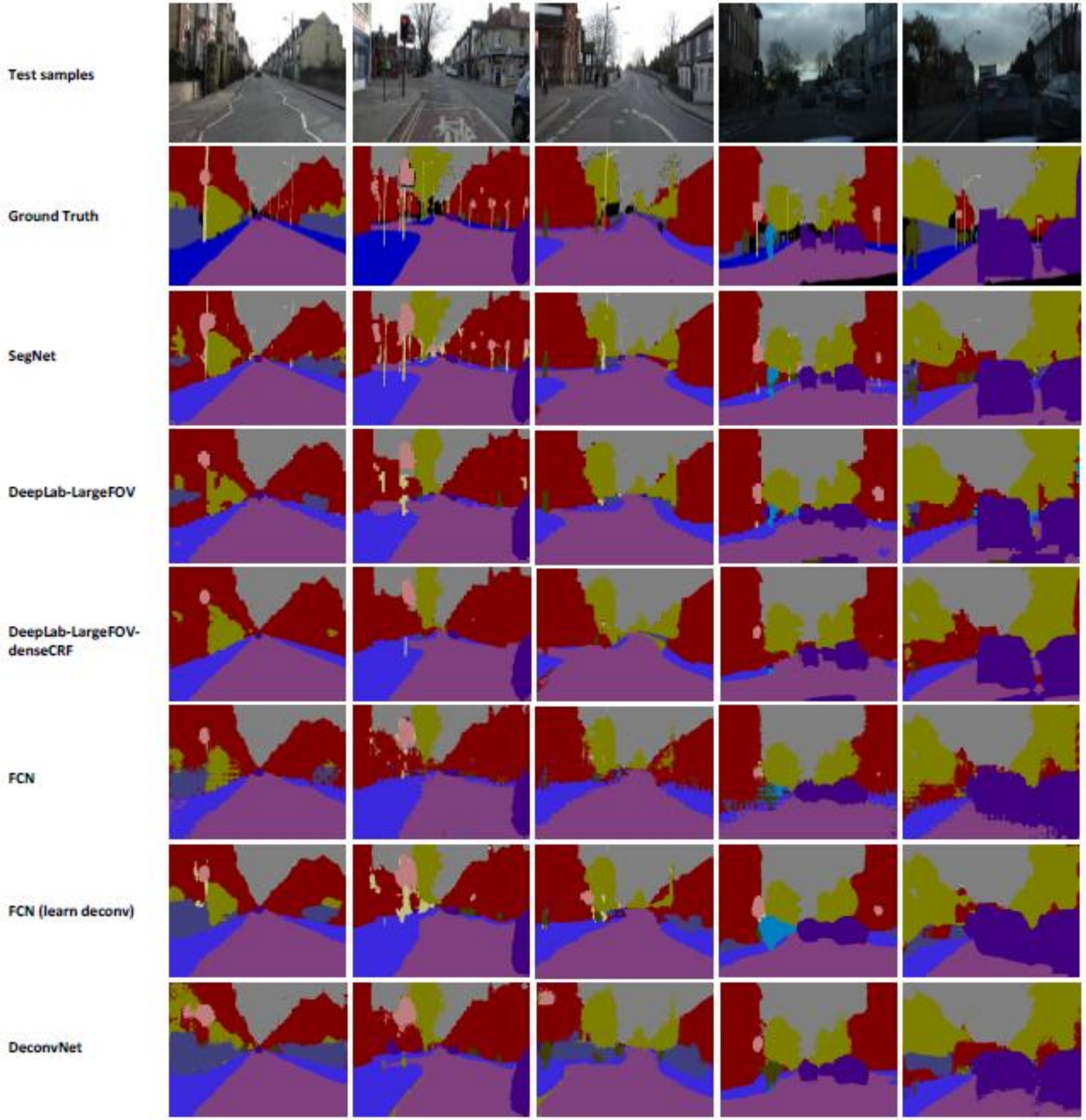


Badrinarayanan et al., 2015

SegNet results

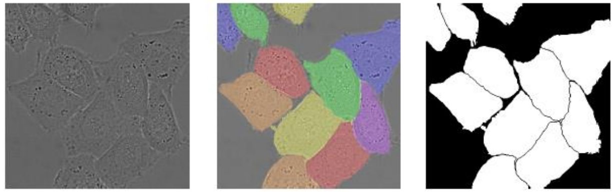


http://mi.eng.cam.ac.uk/projects/segnet/demo.php



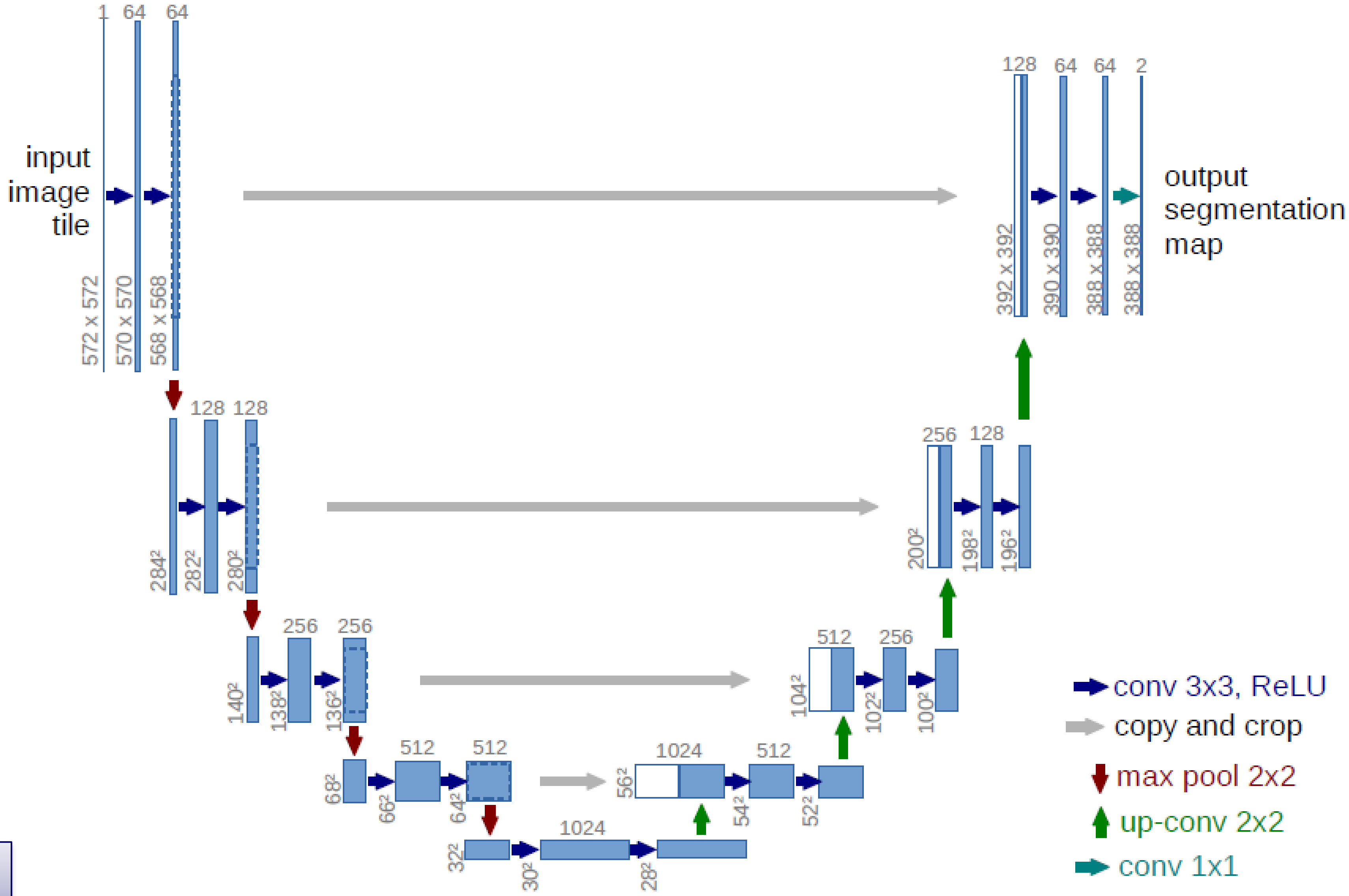
U-net

- Encoder-decoder network
- Contractive and expansive path
- Shortcut connections
- Does not require a lot of training data



Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	**0.9203**	**0.7756**

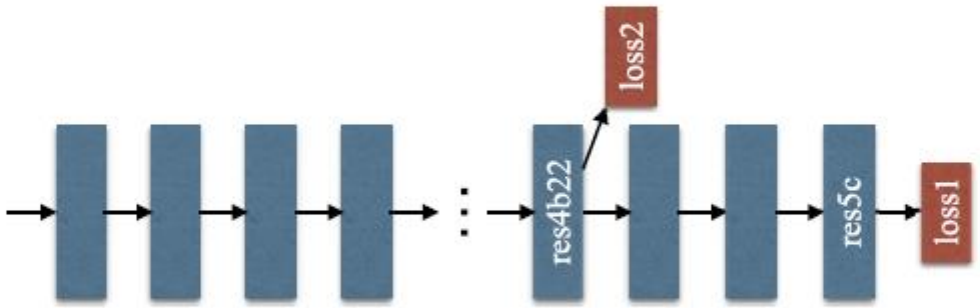
[Ronneberger et al., 2015]

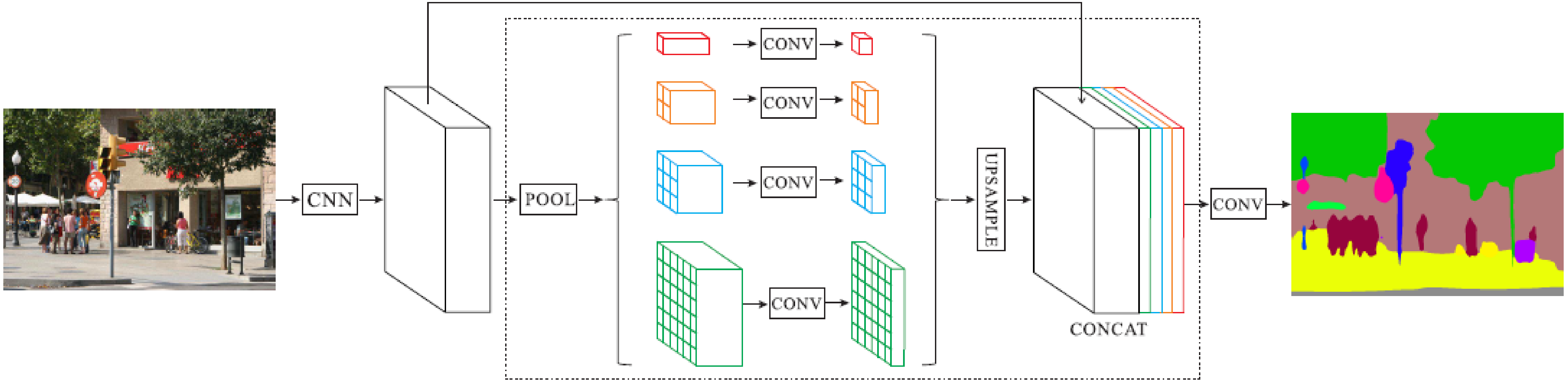


PSP-Net

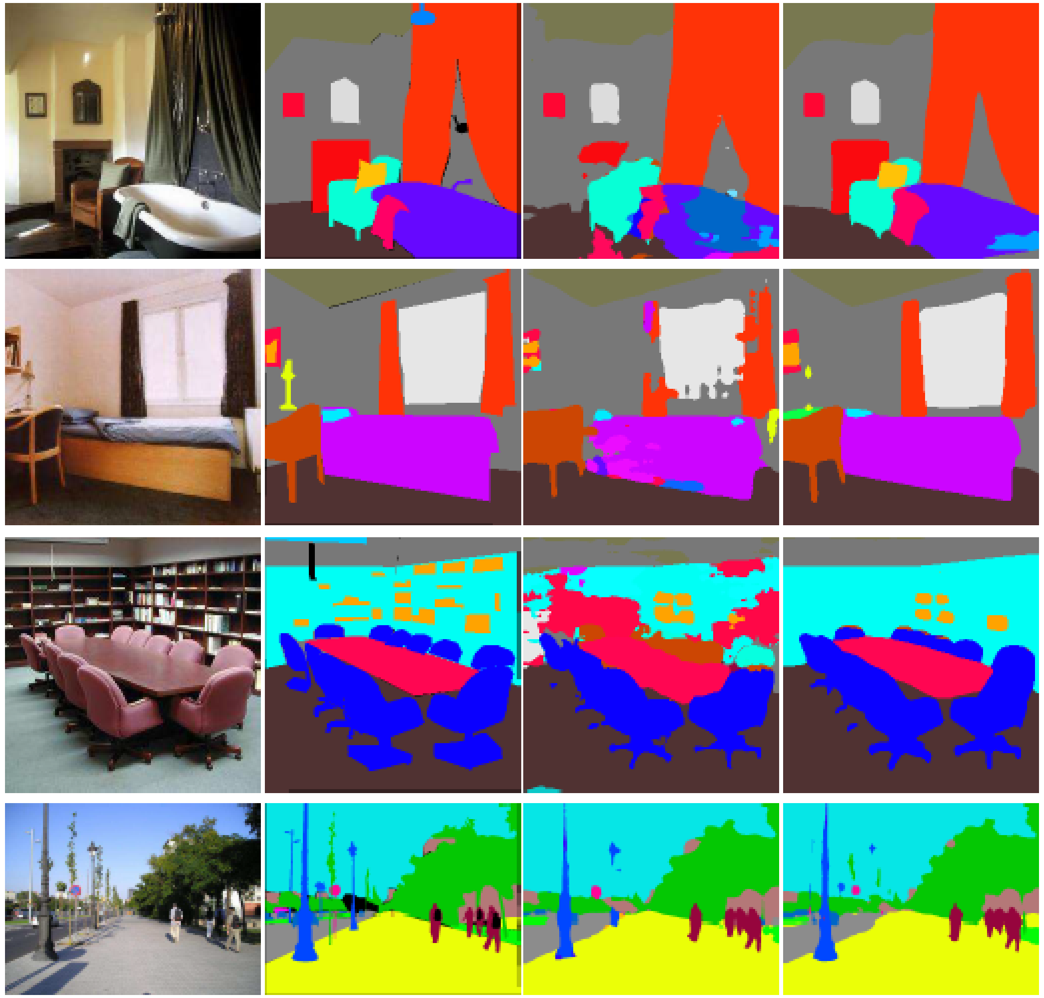
- Pyramid Scene Parsing Network
- Developed for semantic scene segmentation
- ResNet50 backend feature extractor
- Pyramid Pooling Module
- Auxilliary loss

Zhao et al., 2017





PSP-Net

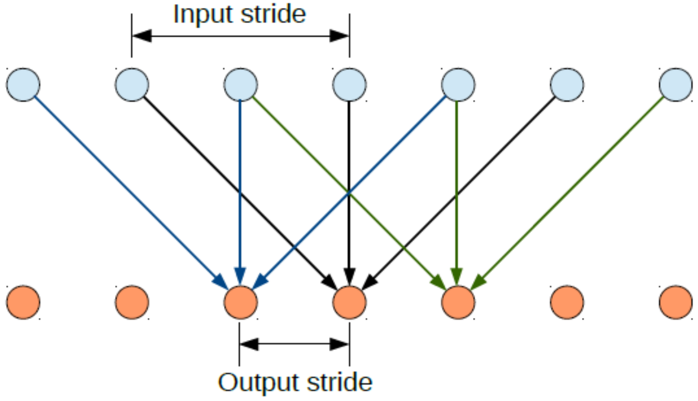


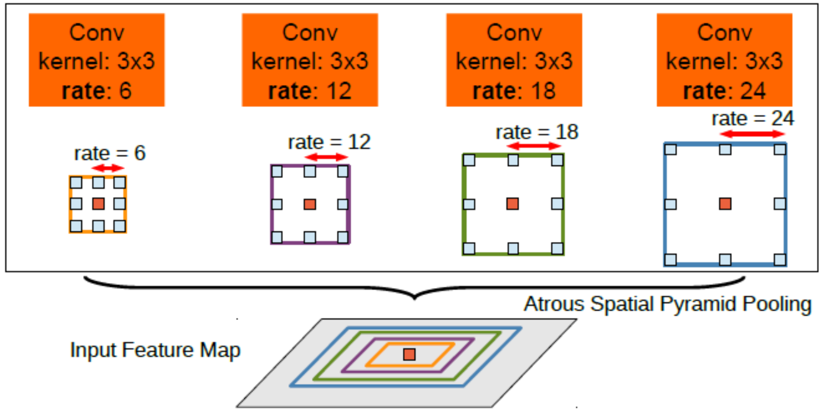
(a) Image (b) Ground Truth (c) Baseline (d) PSPNet

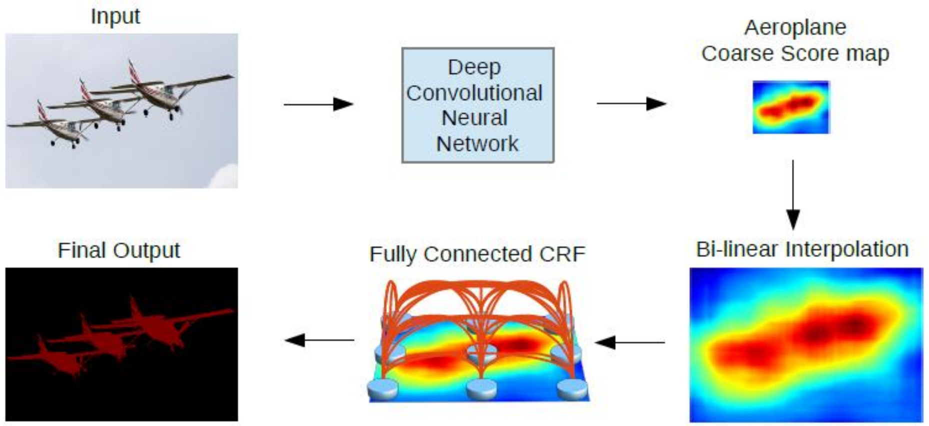
[Zhao et al., 2017]

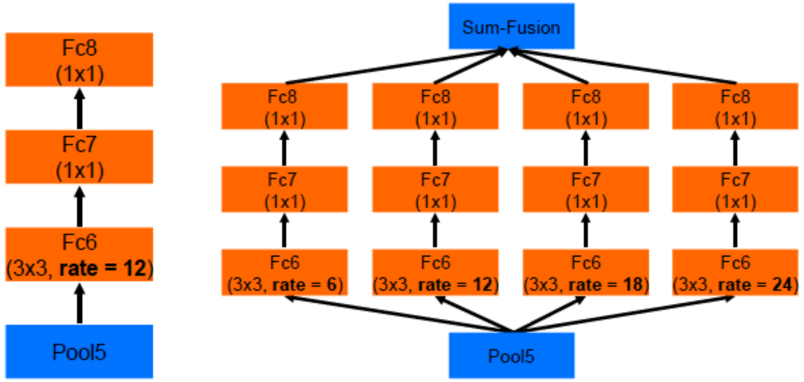
DeepLab

- Based on pretrained VGG-16 (v1) and ResNet101 (v2)
- Atrous convolution
- Fully-connected Conditional Random Fields
- Atrous Spatial Pyramid Pooling
- Multiscale structure
- Cross-entropy loss









(a) DeepLab-LargeFOV (b) DeepLab-ASPP

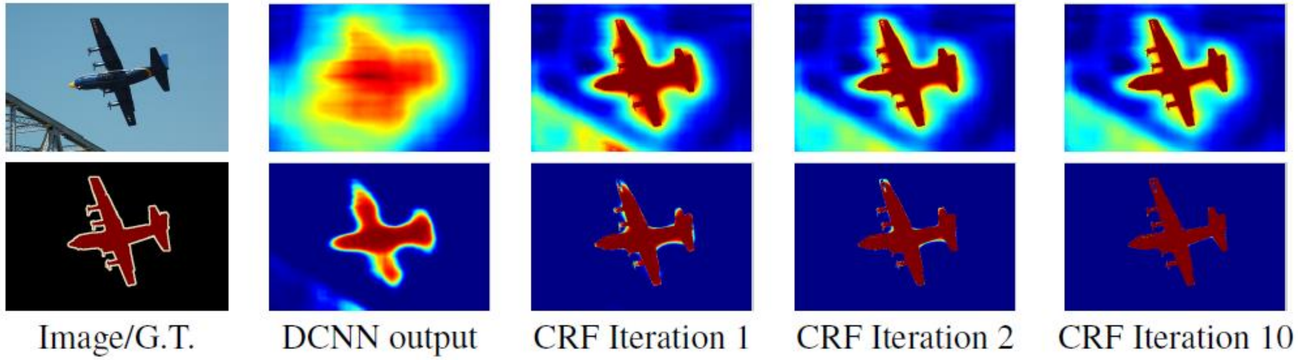
[Zhao et al., 2015]

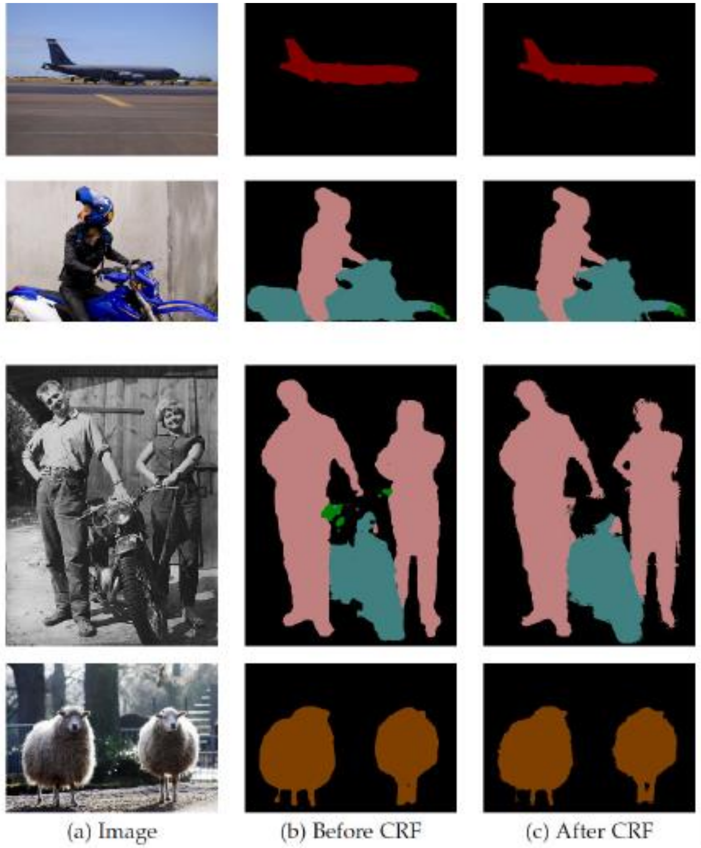
[Zhao et al., 2016]

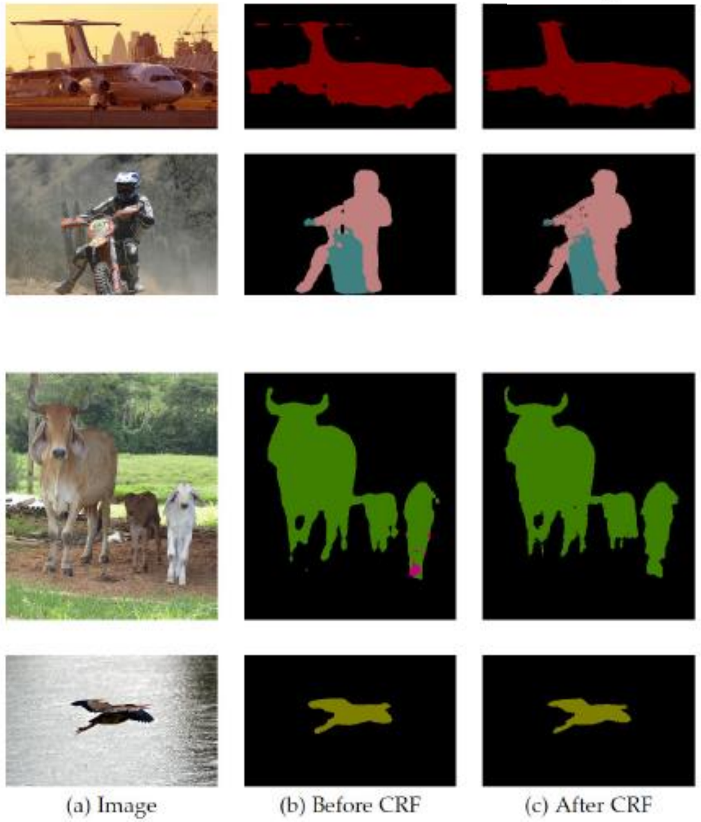
DeepLab

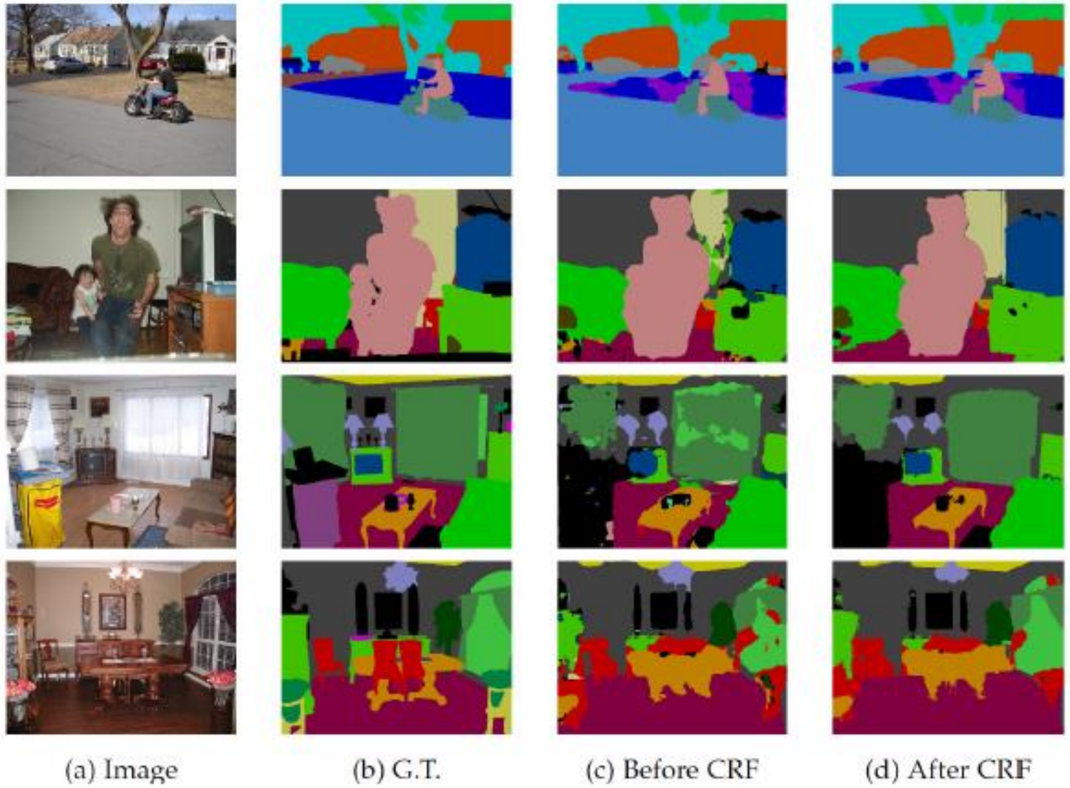
- DeepLabV2 results

Zhao et al., 2016



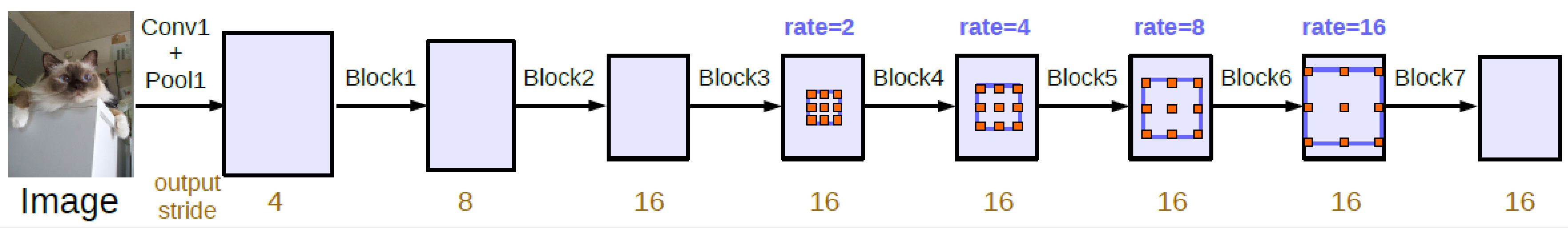


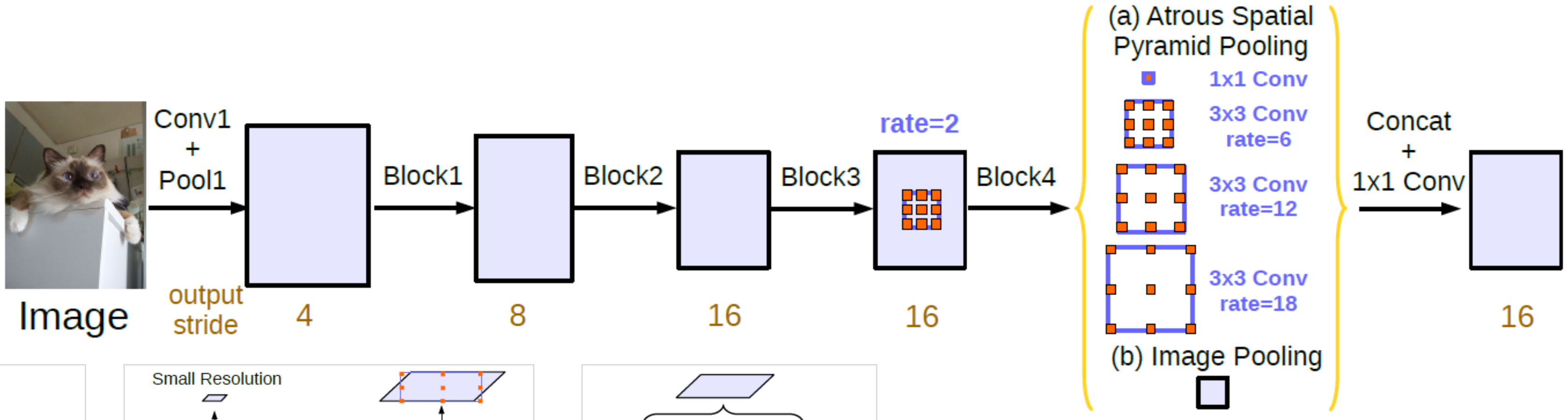


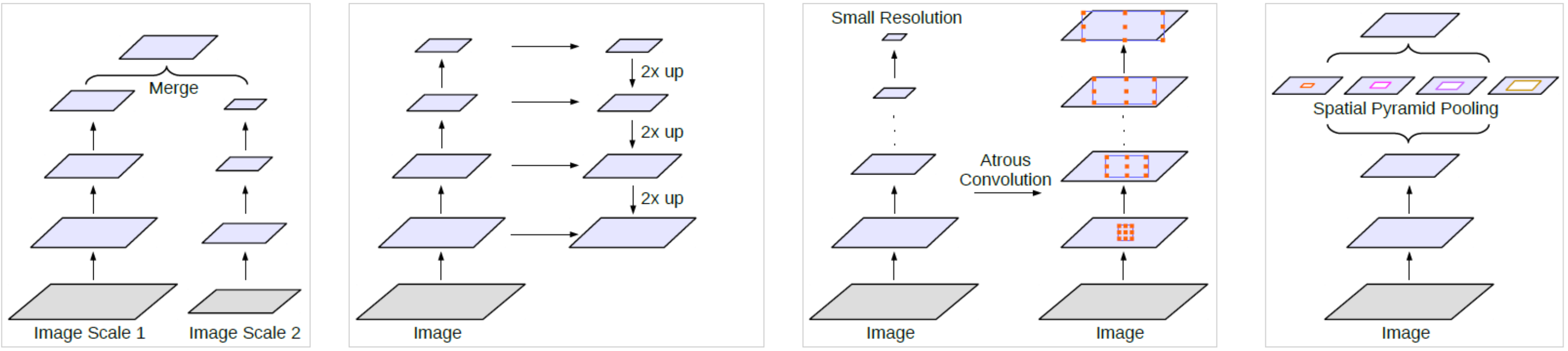


DeepLab v3

- Going deeper with atrous convolutions
- Better ASPP
- Multi-grid, Multi-scale and Output Strides
- ResNet backbone
- Without CRF
- Analysing different architectures



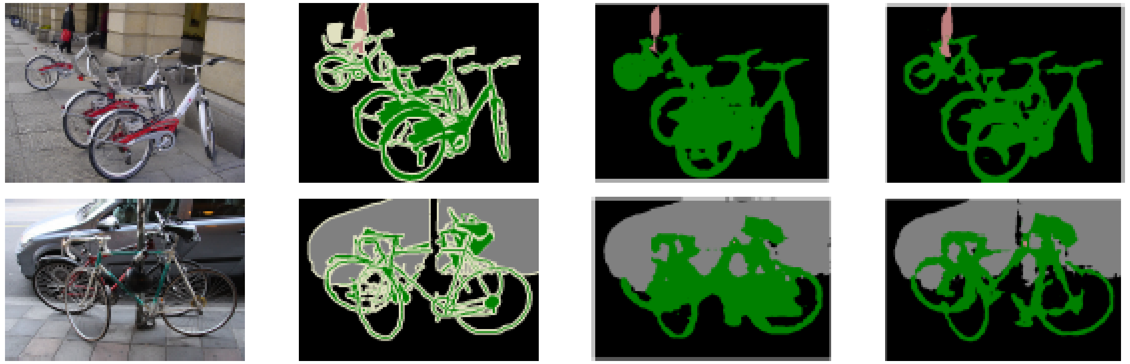




(a) Image Pyramid (b) Encoder-Decoder (c) Deeper w. Atrous Convolution (d) Spatial Pyramid Pooling

[Zhao et al., 2017]

DeepLab v3

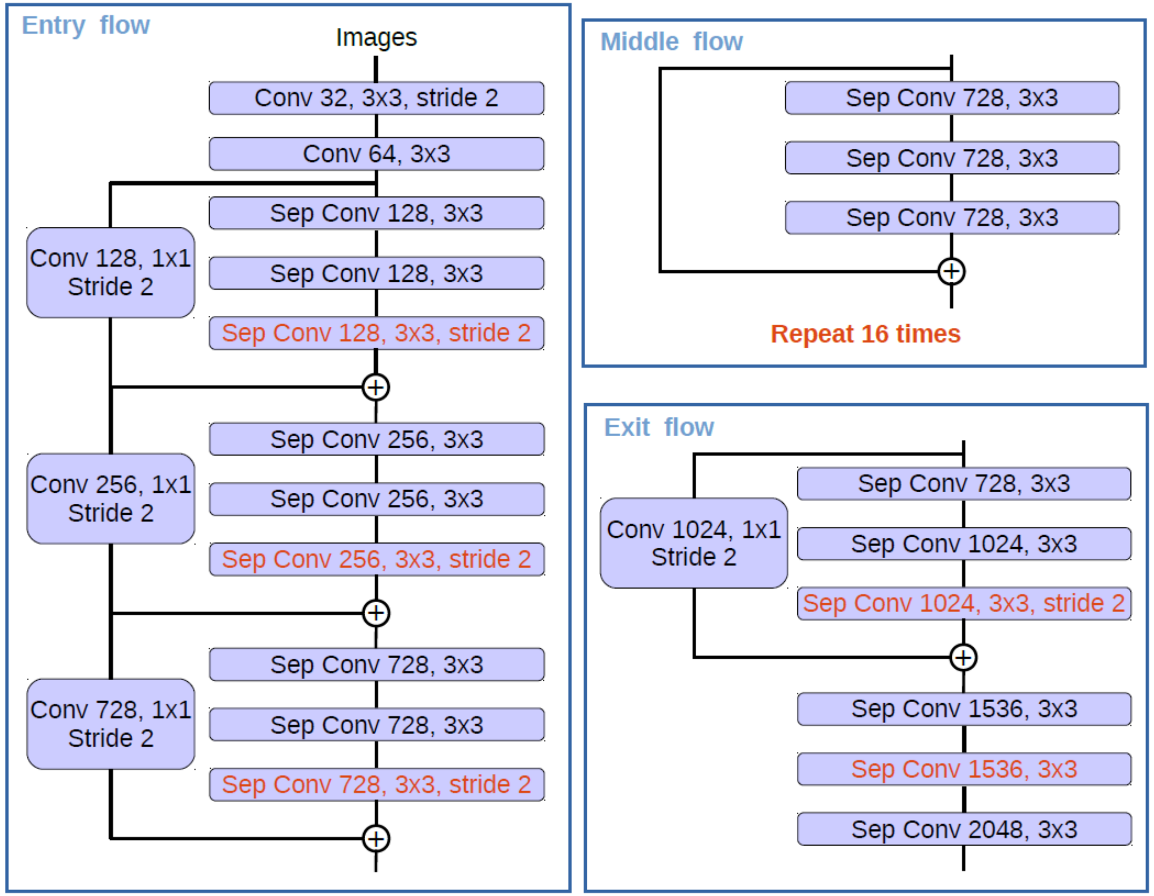


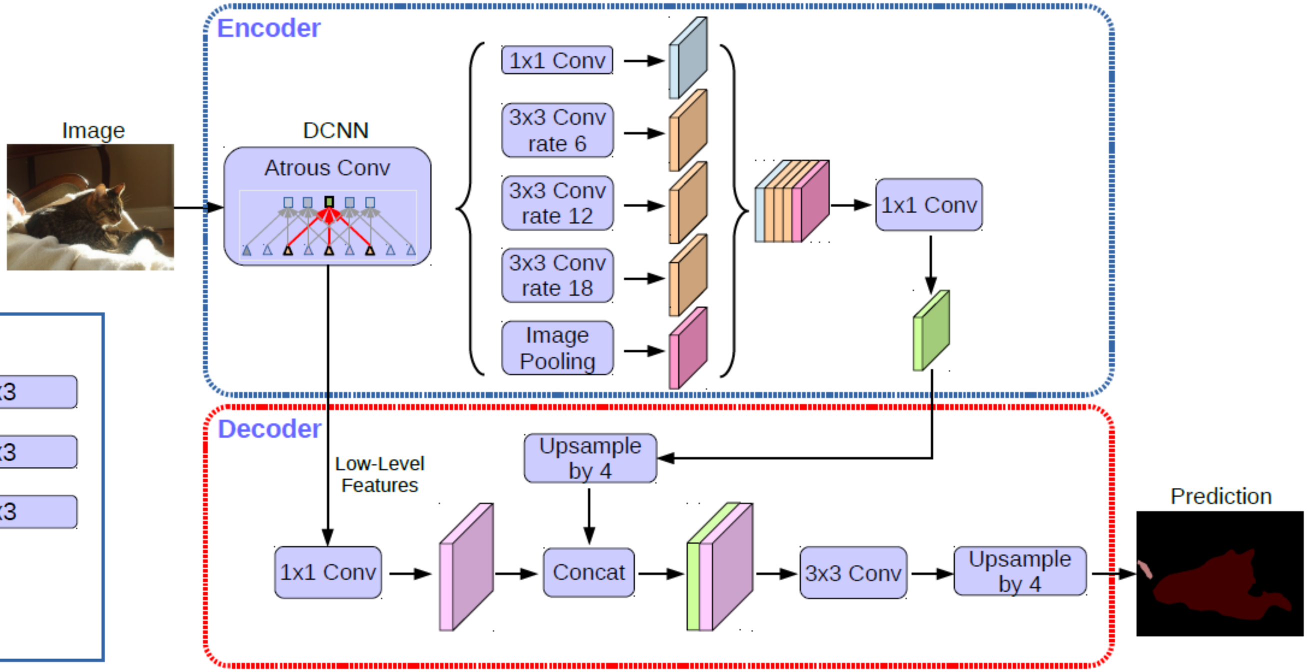
(a) Image (b) G.T. (c) w/o bootstrapping (d) w/ bootstrapping

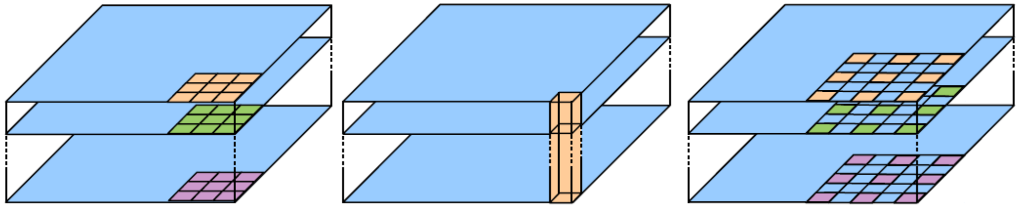
Zhao et al., 2017

DeepLab V3+

- Encoder-decoder architecture
- Atrous depth-wise convolution
- Modified Aligned Xception







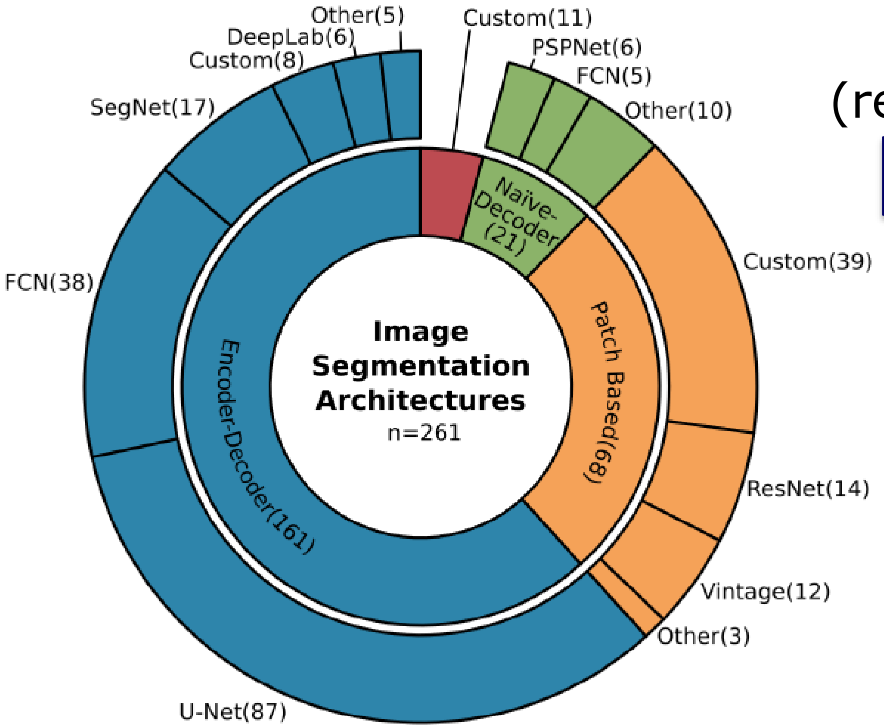
(a) Depthwise conv. (b) Pointwise conv. (c) Atrous depthwise conv.

Zhao et al., 2018

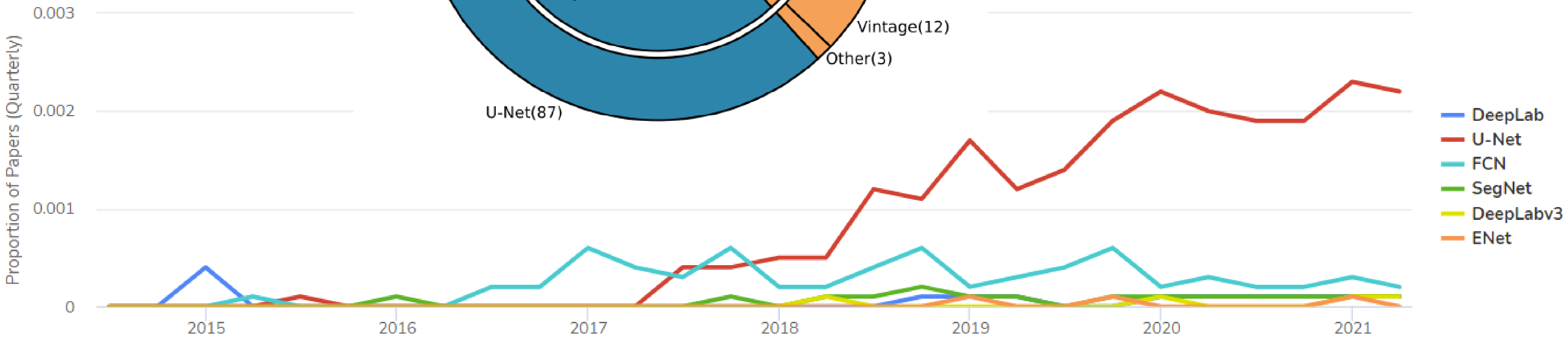
Semantic segmetation arhitectures overview

(remote sensing domain)

[Hoeser et. al 2020]



Usage Over Time



This feature is experimental; we are continuously improving our matching algorithm.

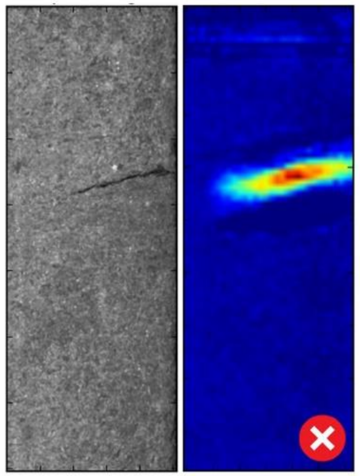
[paperswithcode.com, 2021]

Beyond segmentation

- Image to image translation
- Optical flow estimation
- Depth estimation
- Monocular depth estimation
- Normal estimation
- Edge detection
- Superresolution
- Colouring
- Image enhancement, deblurring
- Surface anomaly detection
- Inpainting
- Counting/density estimation
- Video segmentation
- Image restoration
- Image synthesis,…



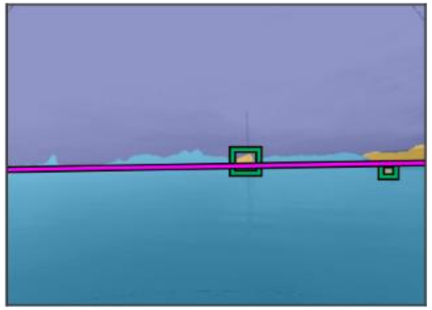
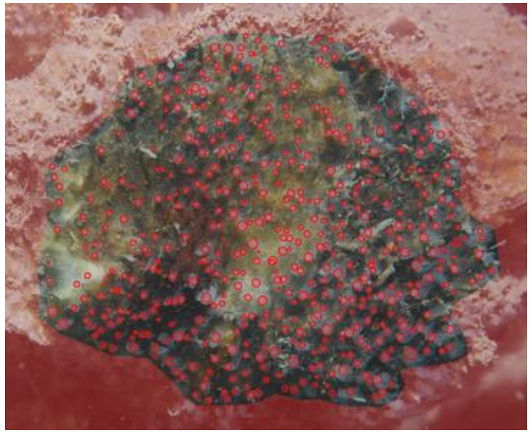
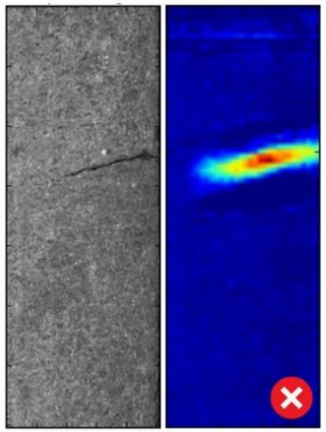


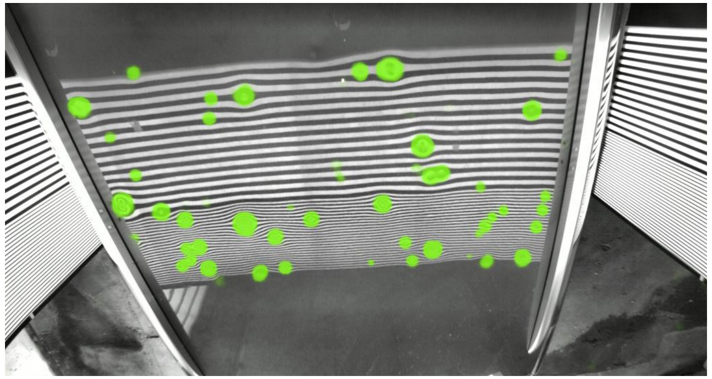
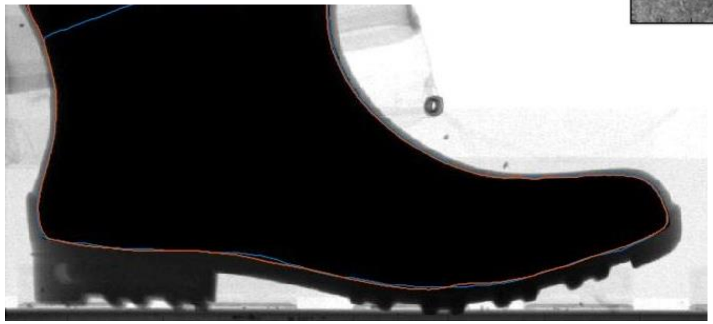
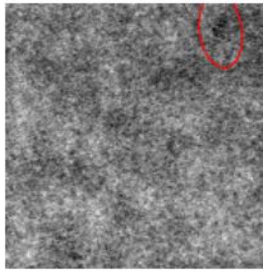



Liu et al., Zhang et al., Jonschkowski et. al, Pan et al., ECCV2020

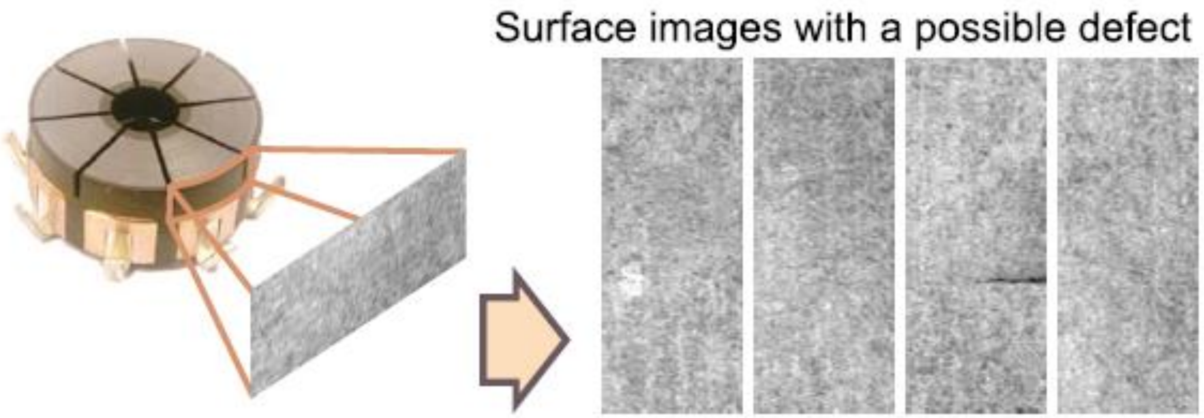
Segmentation for various computer vision tasks

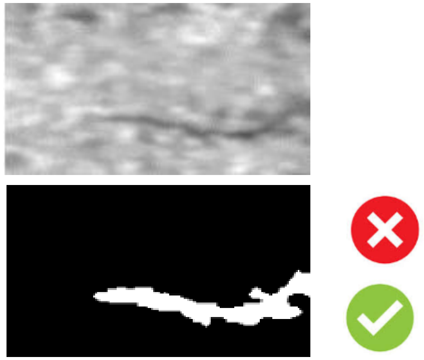
- Detection of surface visual defects
  - industrial products
  - damage on car body
- Polyp counting
- Obstacle detection
- Image enhancement
- Semantic edge detection

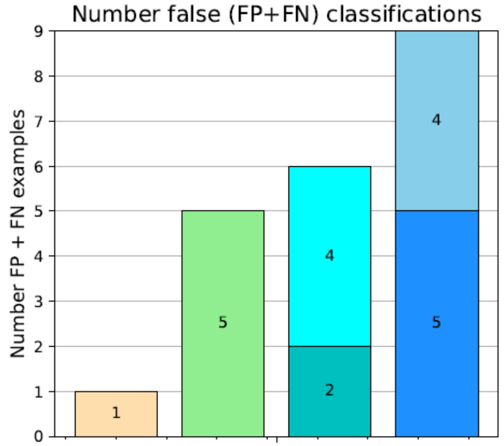




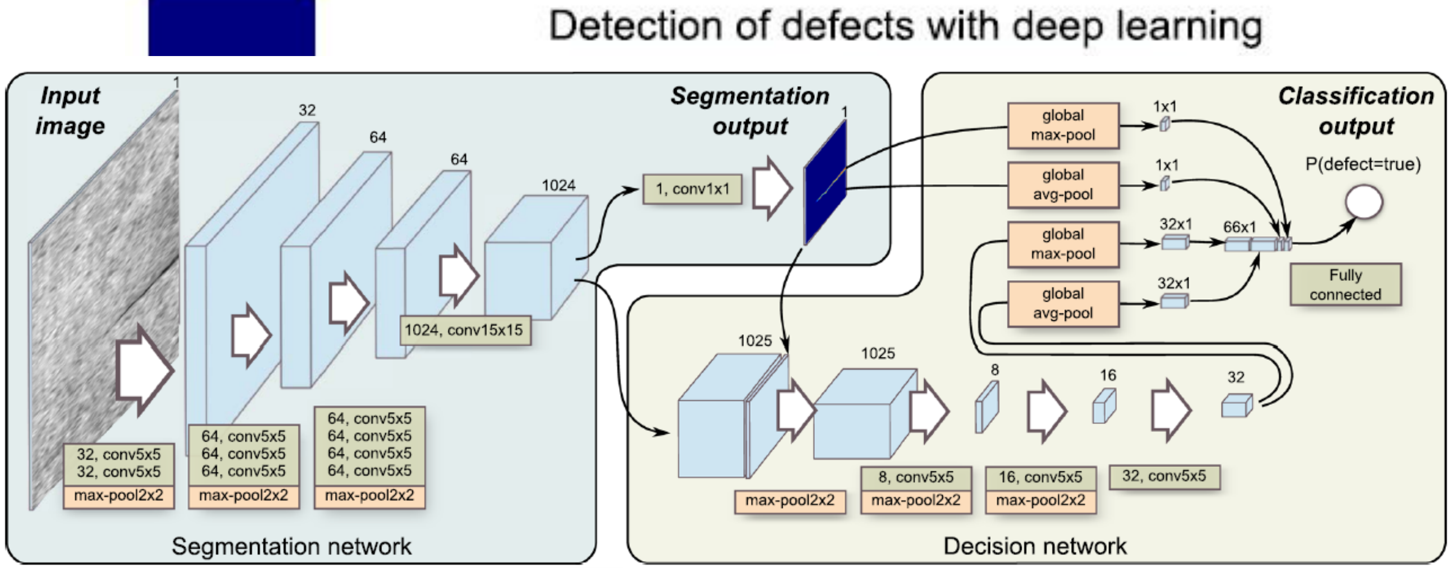




Segmentation-based surface-defect detection

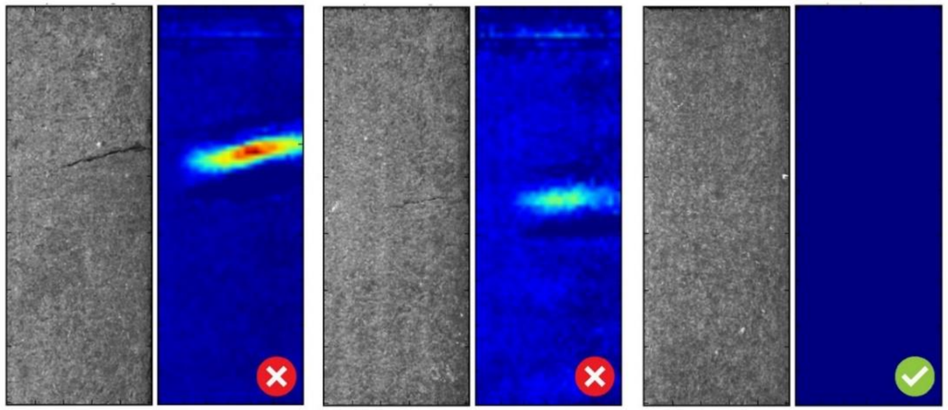








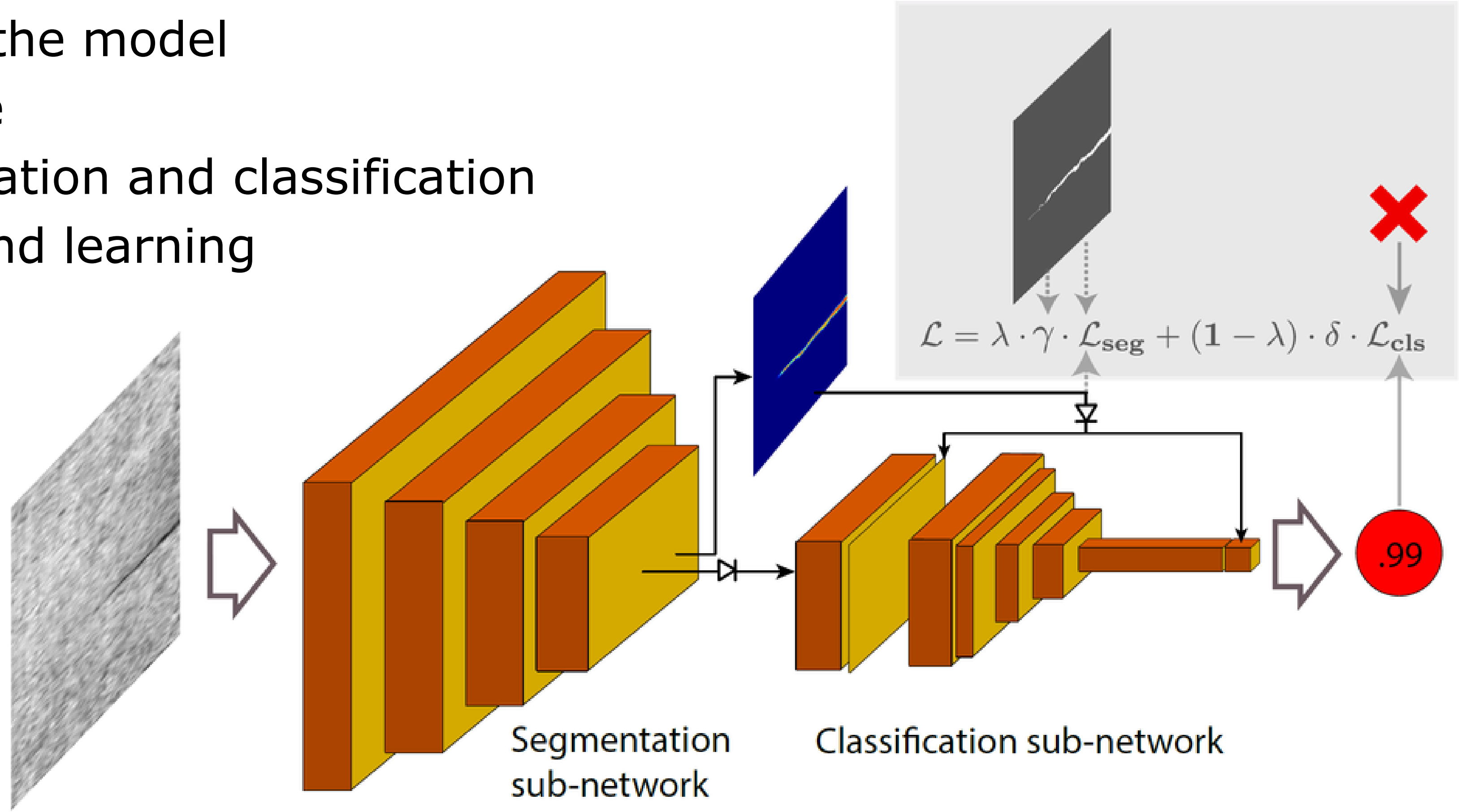
Segmantation-based data-driven surface-defect detection



Tabernik et. al, 2020

End-to-end network architecture

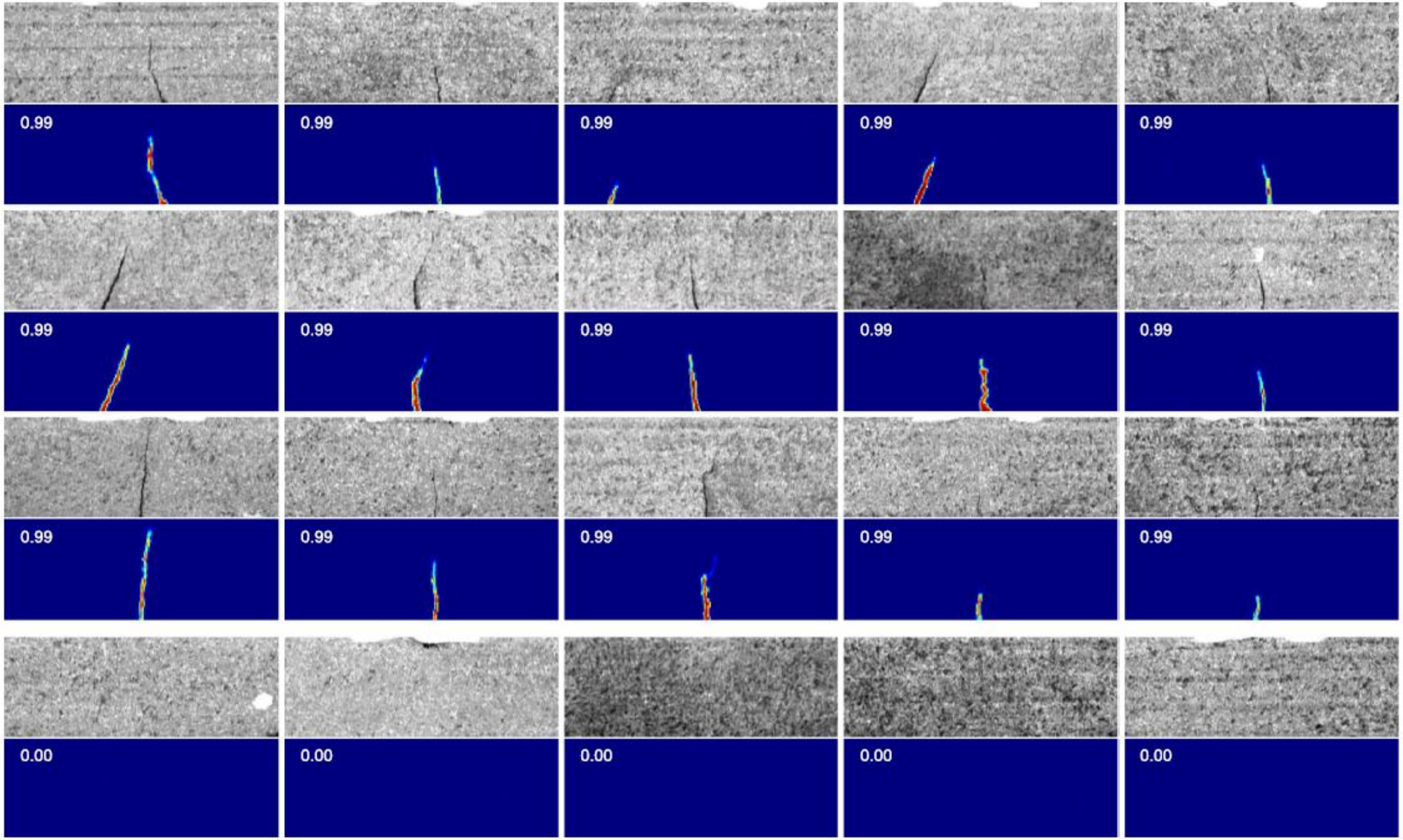
- Training the model
- Inference
- Segmentation and classification
- End-to-end learning



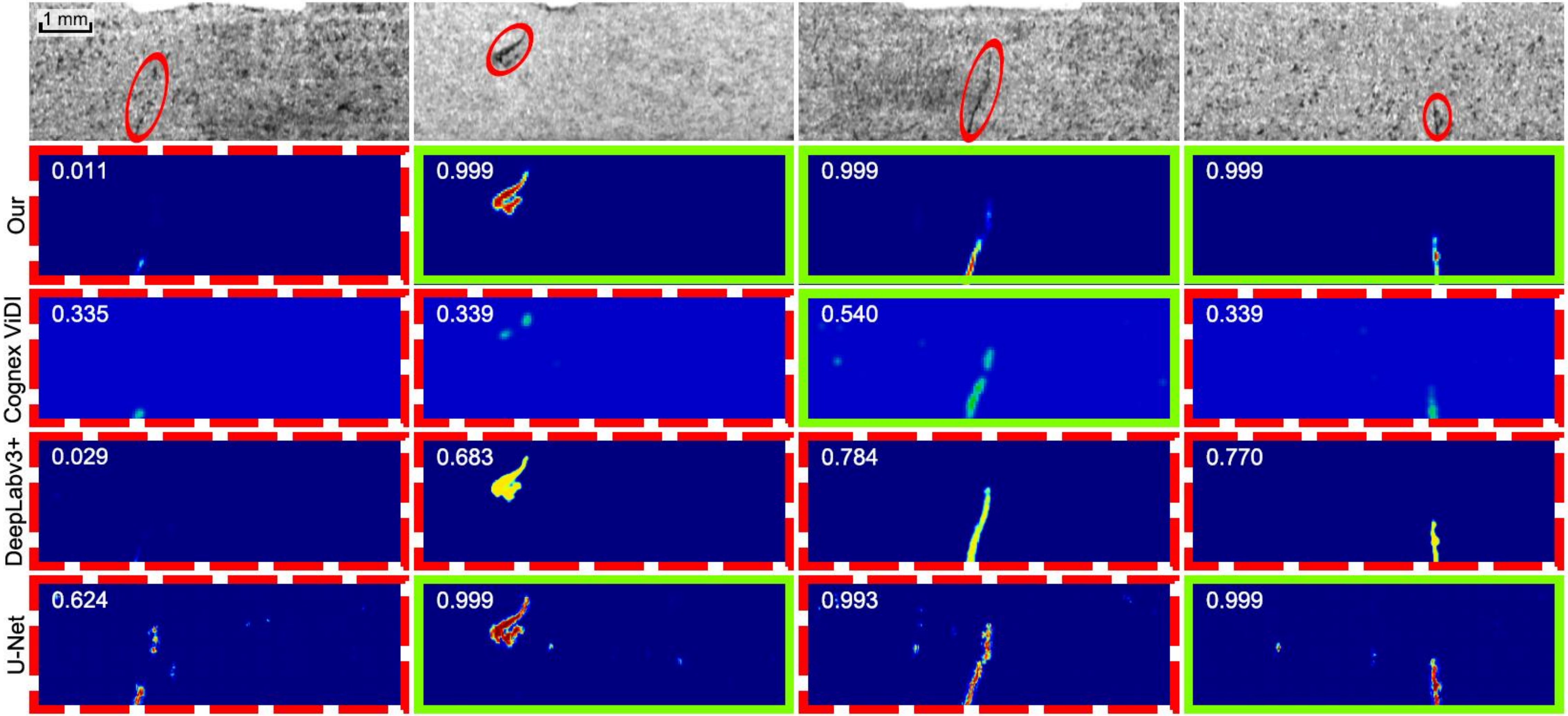
Architecture and approach	*Learning stages*	*Number of positive training samples*					
		33	*25*	*20*	*15*	*10*	*5*
Extended Segmentation+Decision Network (ours)	*end-to-end*	**100.00**	**99.78**	**100.00**	**99.88**	**99.31**	**96.71**
Segmentation+Decision Network [9]	*separate (two stages)*	99.0	97.5	99.5	97.4	98.8	95.8
Cognex ViDi (commercial software) [9]	-	99.0	97.4	95.7	97.1	95.6	89.2

Božič et. al, 2021

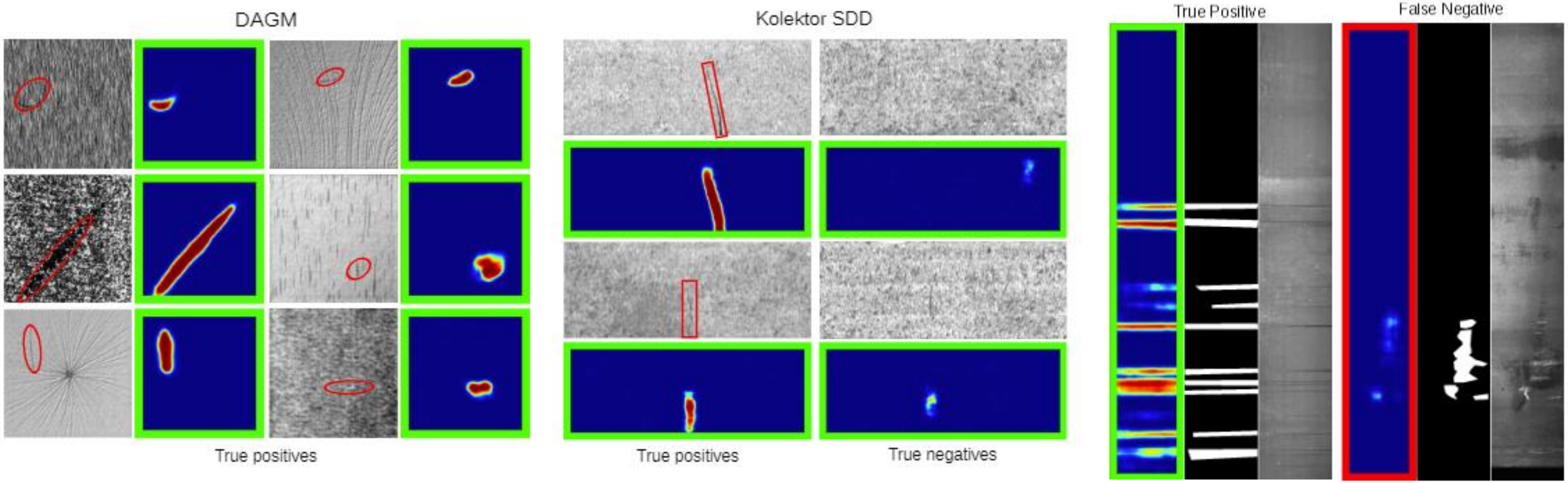
Surface-defect detection

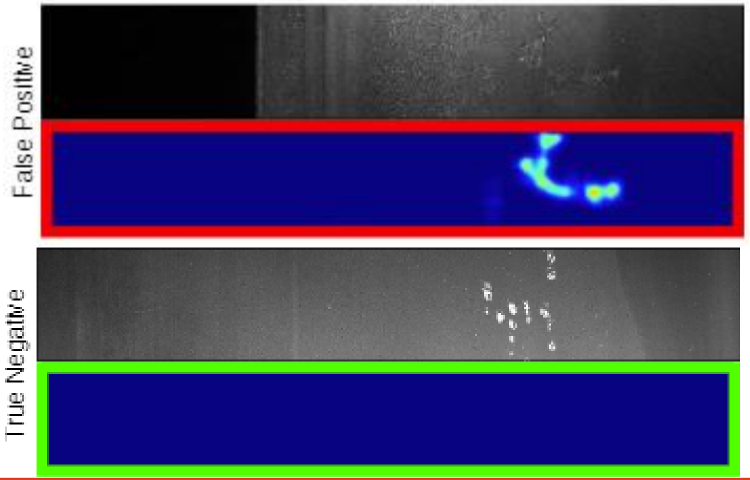


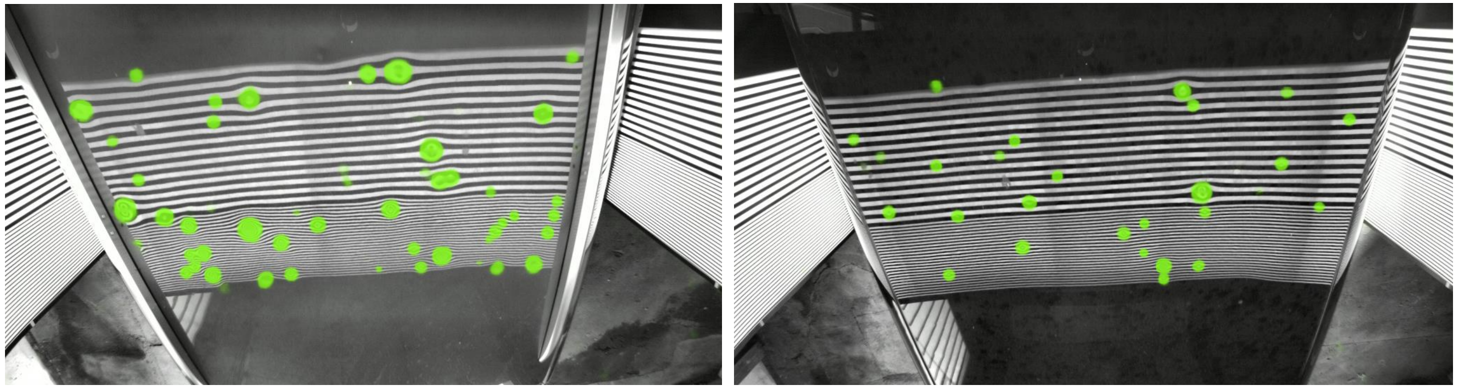
Surface-defect detection



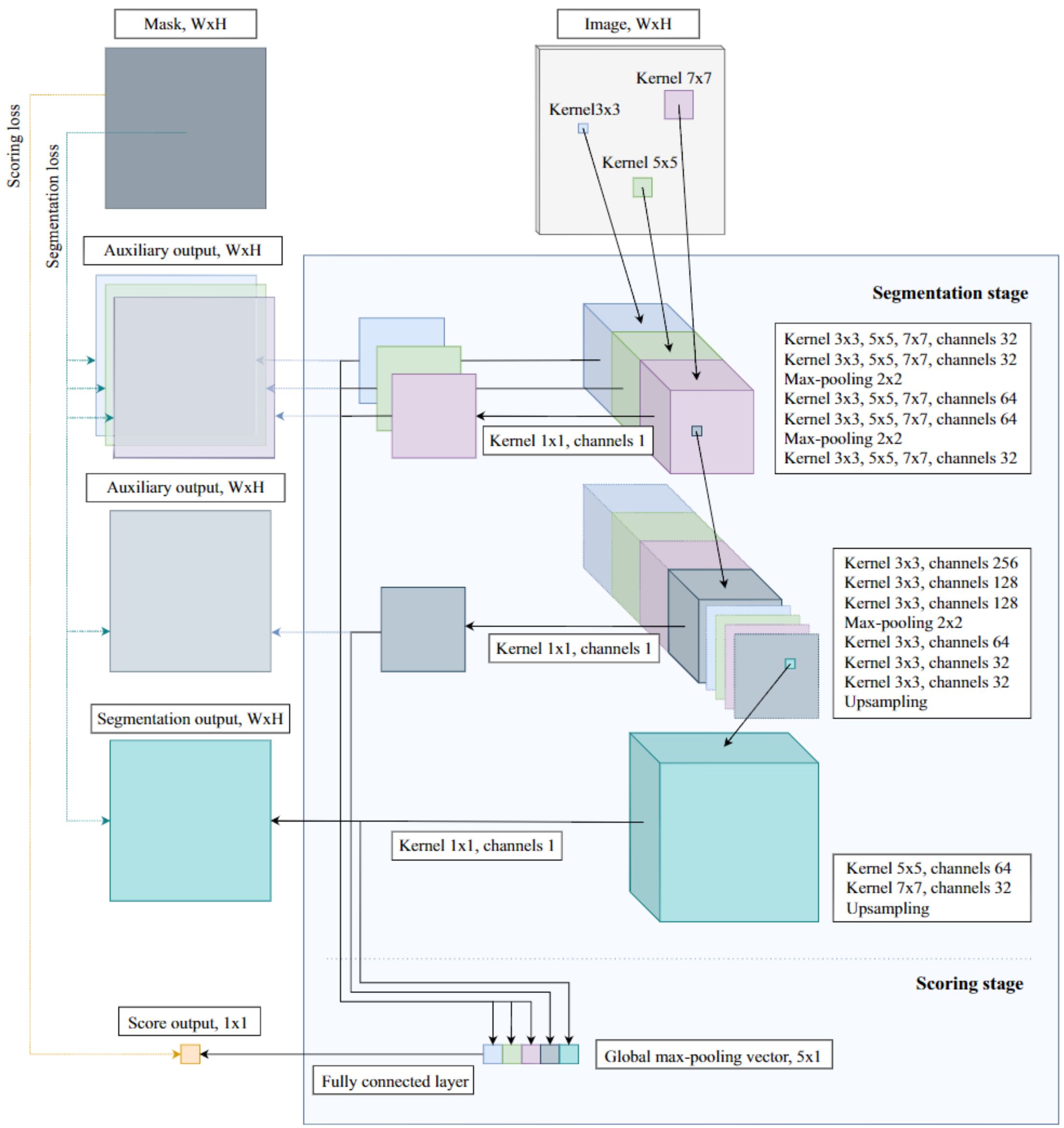
Surface-defect detection

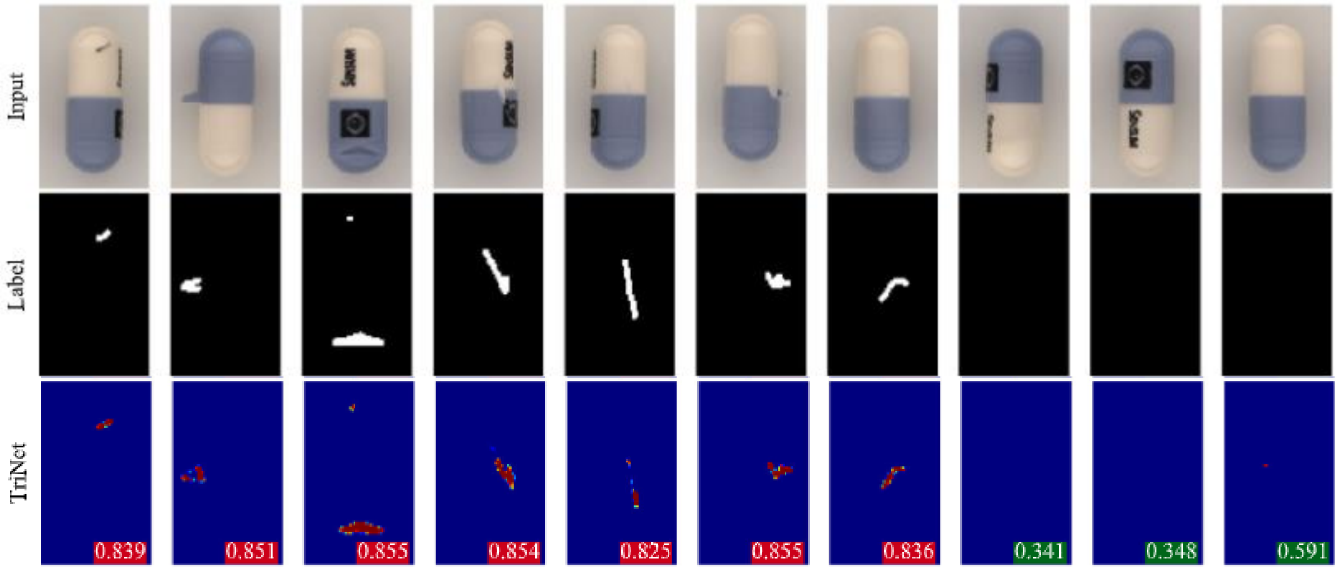


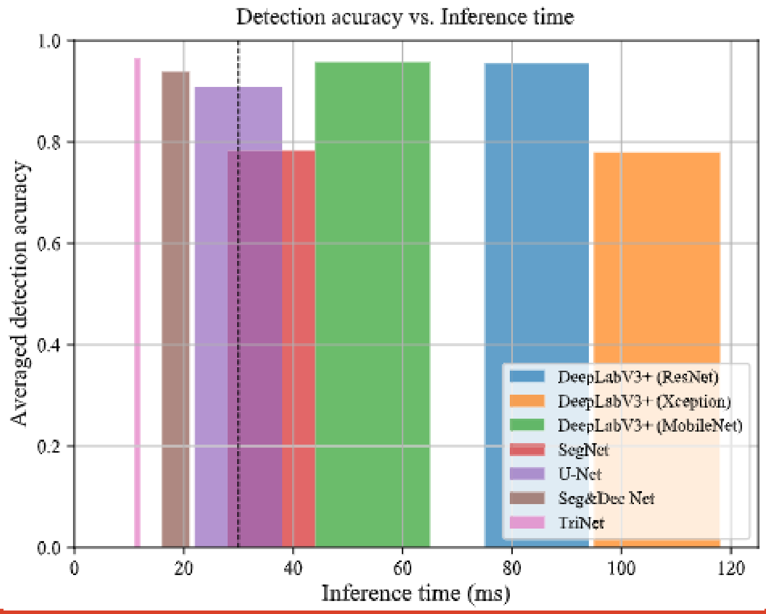


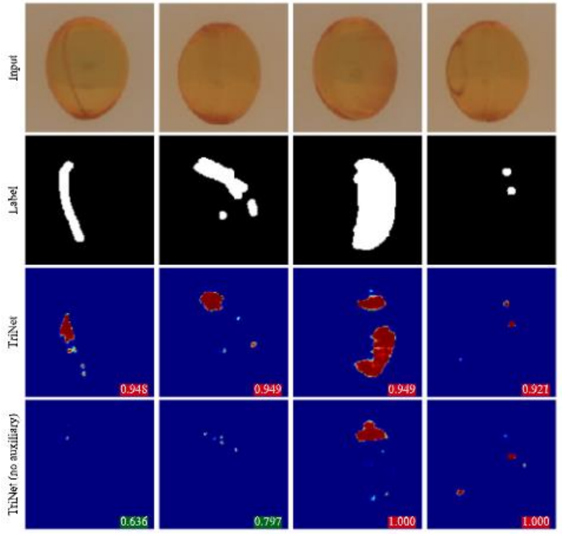


Surface-defect detection





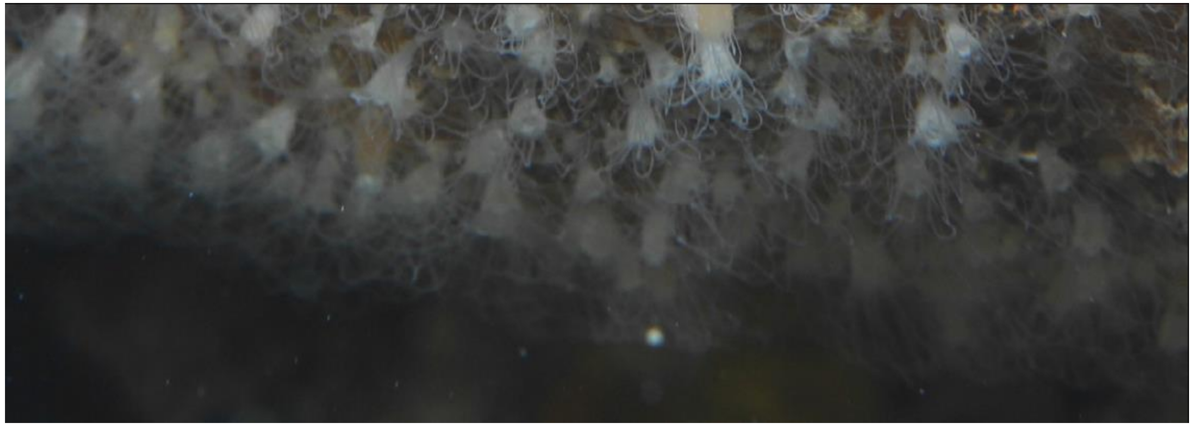


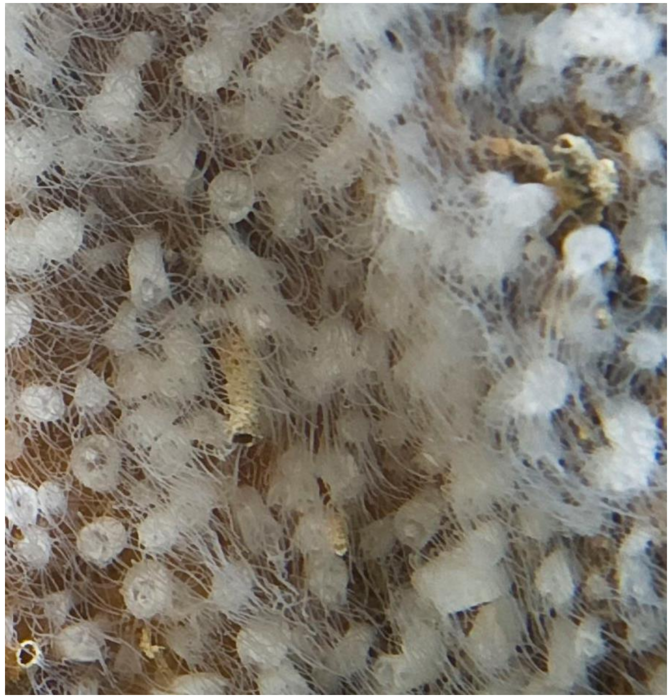


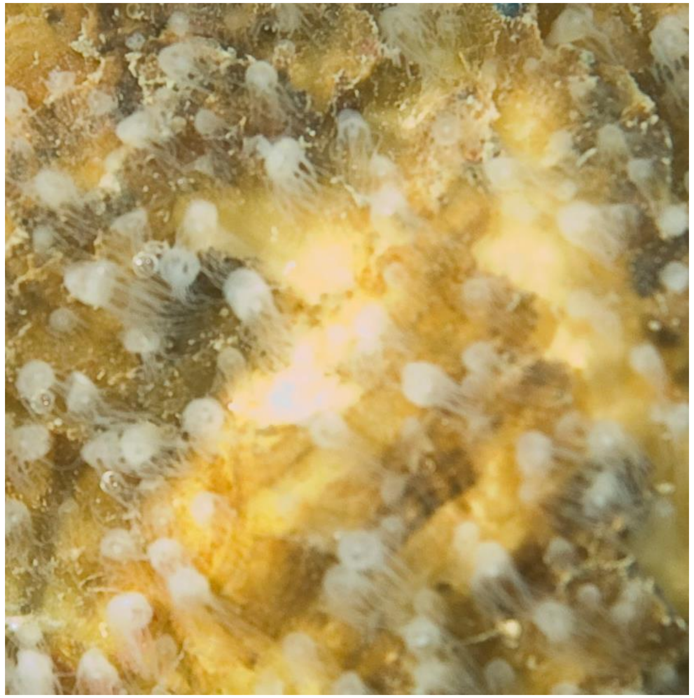
Rački et. al, 2021

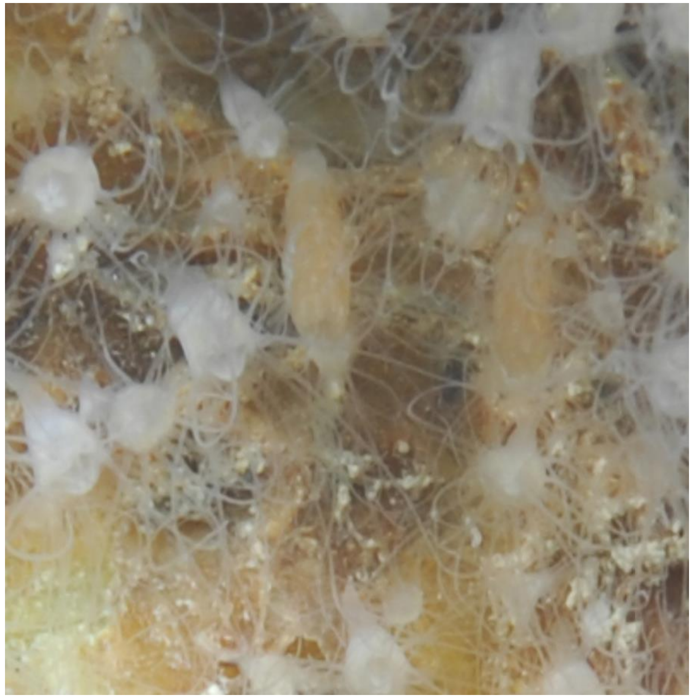
Segmentation for polyp counting

- Segmentation based counting
- Challenges:
  - Appearance variability
  - Blurring
  - Heavy occlusions

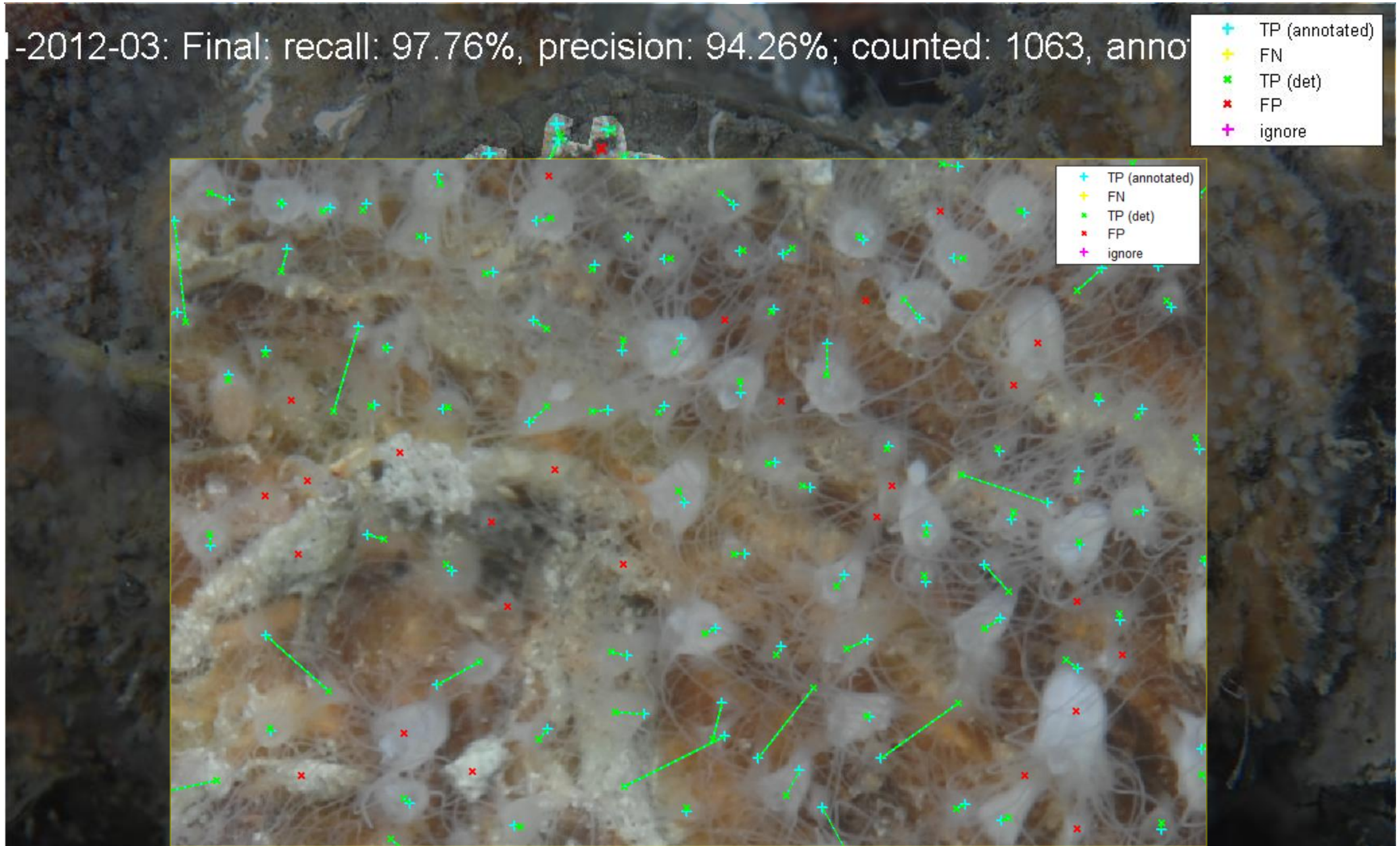








Polyp counting

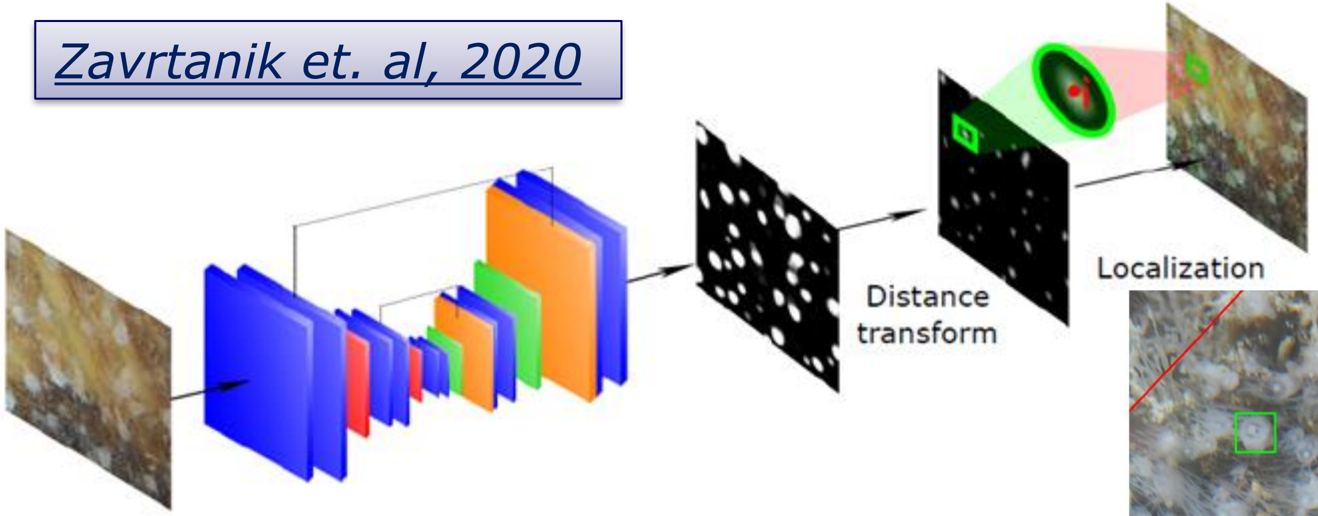


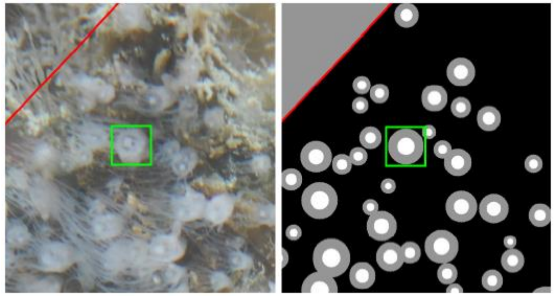
Vodopivec, Mandeljc, Makovec, Malej, Kristan, Polyp counting made easy: towards automated scyphistoma census in underwater imagery

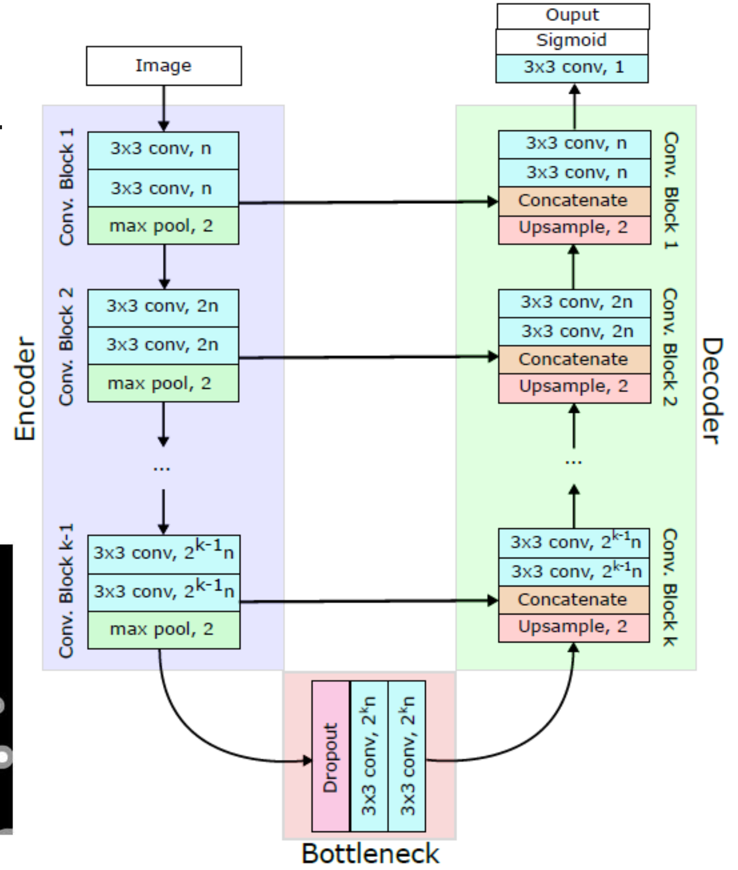
Polyp counting

- U-Net-based architecture for segmentation

[Zavrtanik et. al, 2020]

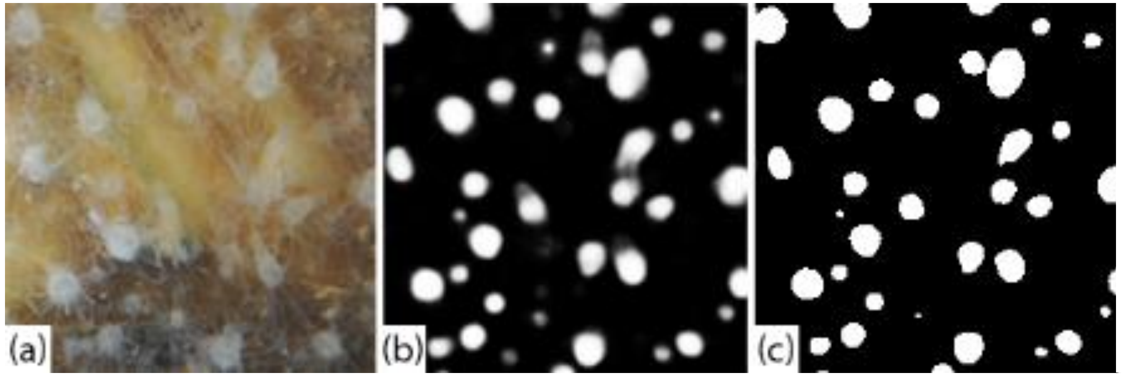






- Thresholding + postprocessing segmentation output:

Input image CNN seg $p > \theta$



Polyp counting

- Data set (37+ 7 images (488x2844), ~50k polyps
  - 7 annotators, ambiguous annotations

Image	Expert diver	Expert annotator	Volunteer	Ground truth	Relative error (max.)
#1	358	378	397	455	17 %
#2	617	571	561	655	14 %
#3	455	453	462	543	17 %
#4	637	678	715	770	17 %
#5	622	676	744	723	14 %
#6	336	296	270	350	23 %
#7	384	304	323	398	24 %

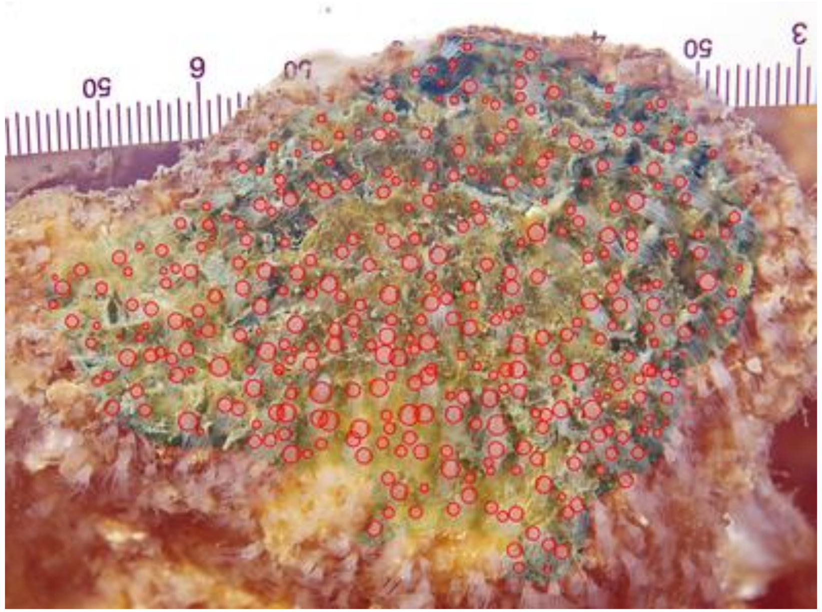


Image	Volunteer 1			
	Day 1	Day 2	Day 3	Day 4
#5	490	472	576	597

Method	Ratio	Rel. err.	AP	AR	*F-1*
SegCo$^{(4,64)}$	**0.99 ± 0.02**	**0.01 ± 0.02**	0.95 ± 0.02	**0.94 ± 0.01**	**0.94 ± 0.01**
SegCo$^{(4,16)}$	0.96 ± 0.03	0.04 ± 0.03	**0.96 ± 0.02**	0.92 ± 0.03	**0.94 ± 0.01**
PoCo Vodopivec et al. (2018)	0.82 ± 0.16	0.23 ± 0.08	0.79 ± 0.08	0.63 ± 0.06	0.70 ± 0.03
RetinaNet	0.92 ± 0.05	0.08 ± 0.05	**0.96 ± 0.02**	0.89 ± 0.04	0.92 ± 0.01

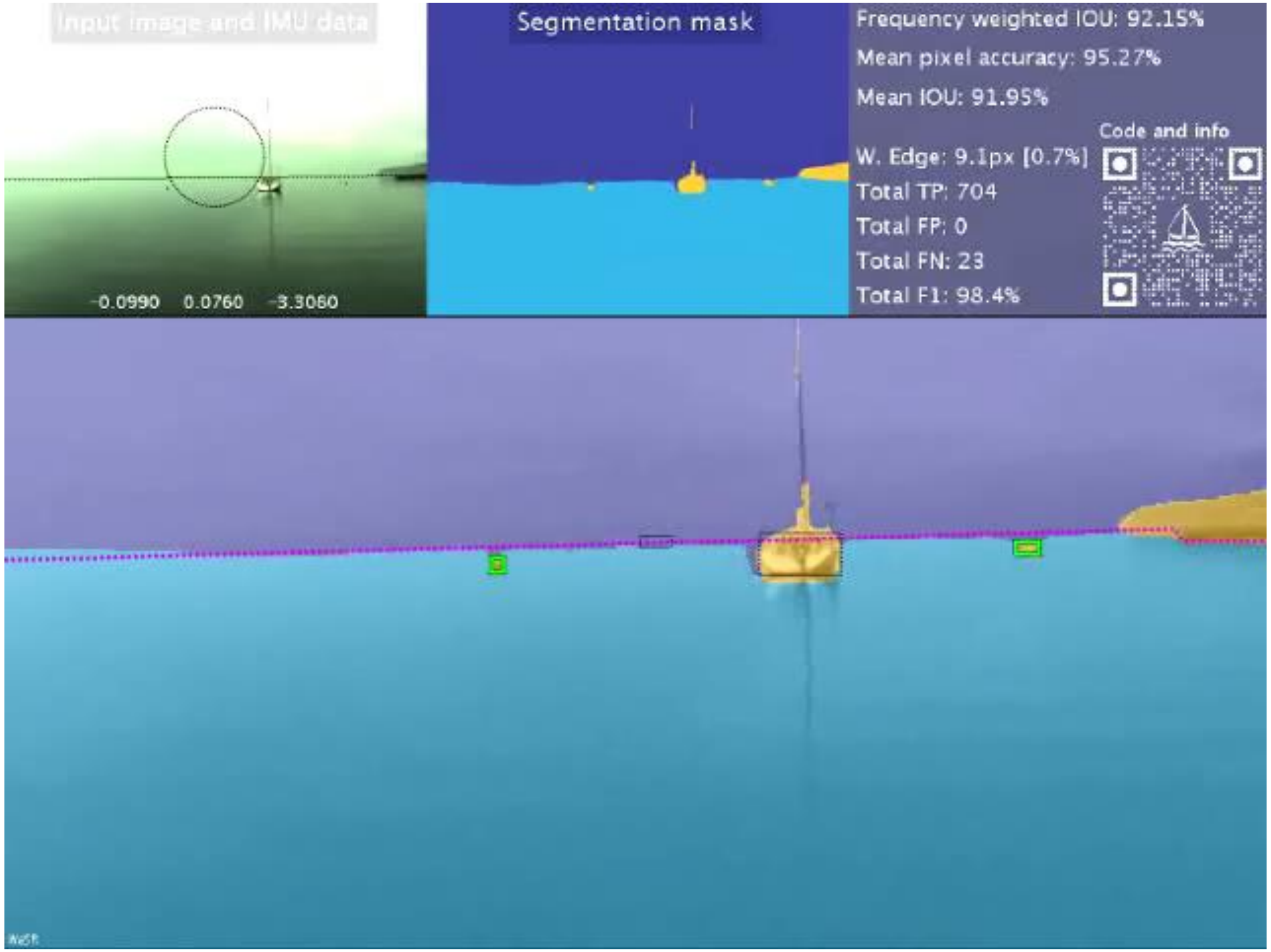
Semantic segmentation for obstacle detection

Bovcon & Kristan, 2020

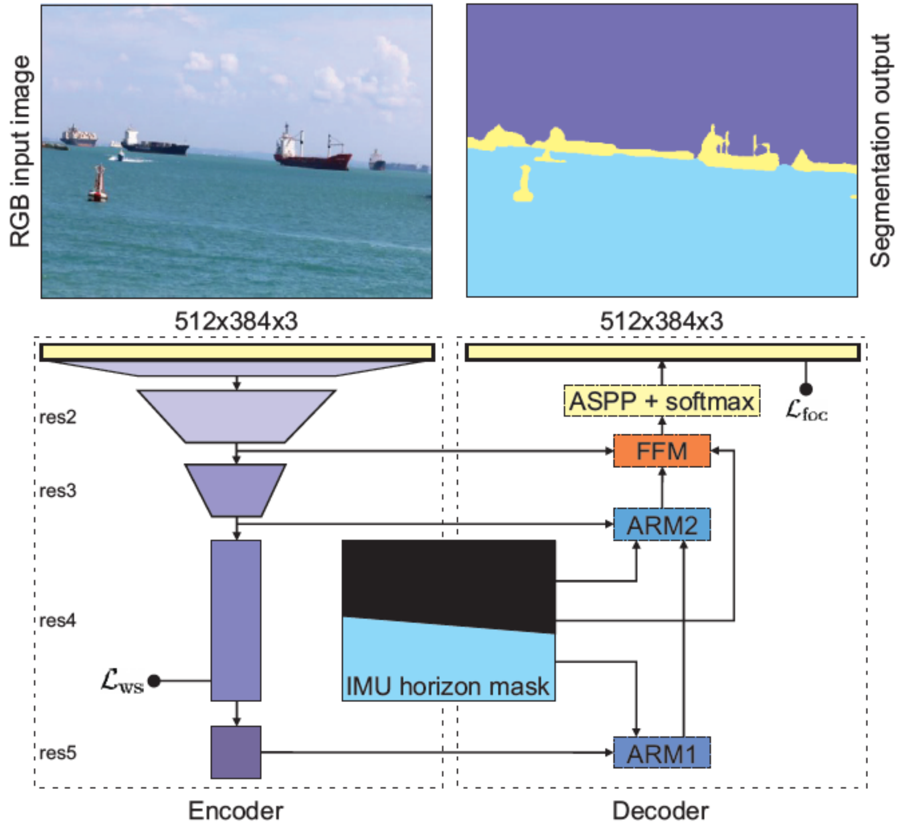
USV equipped with different sensors:

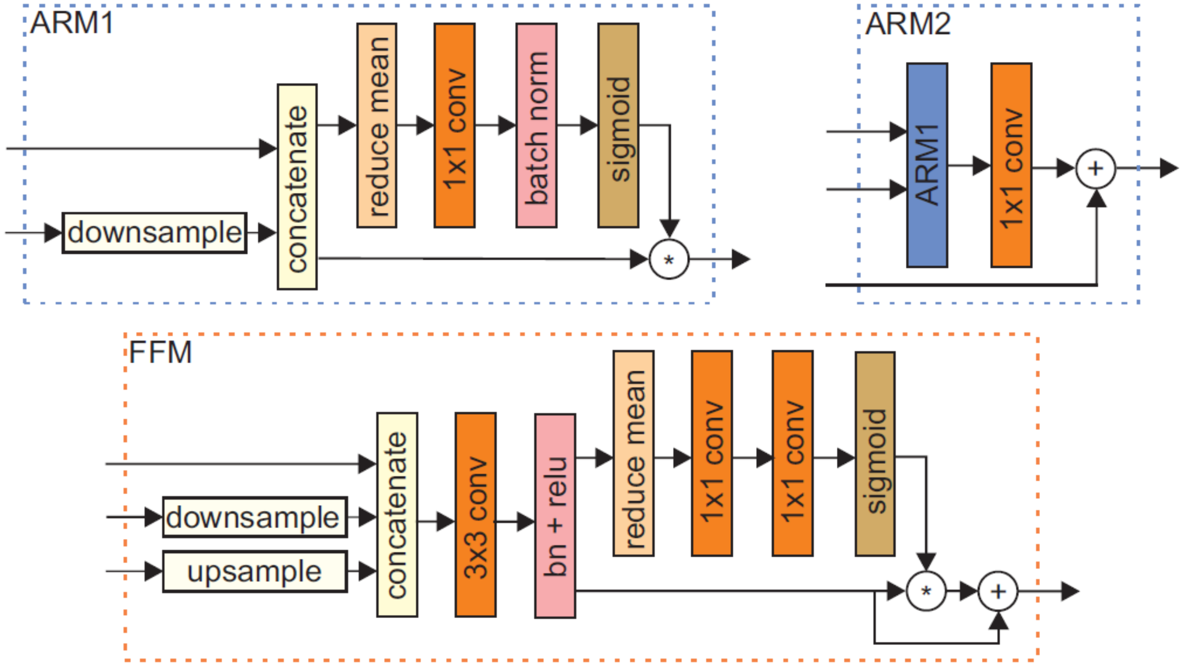
- stereo camera
- IMU
- GPS
- compass

Segmentation based on RGB + IMU

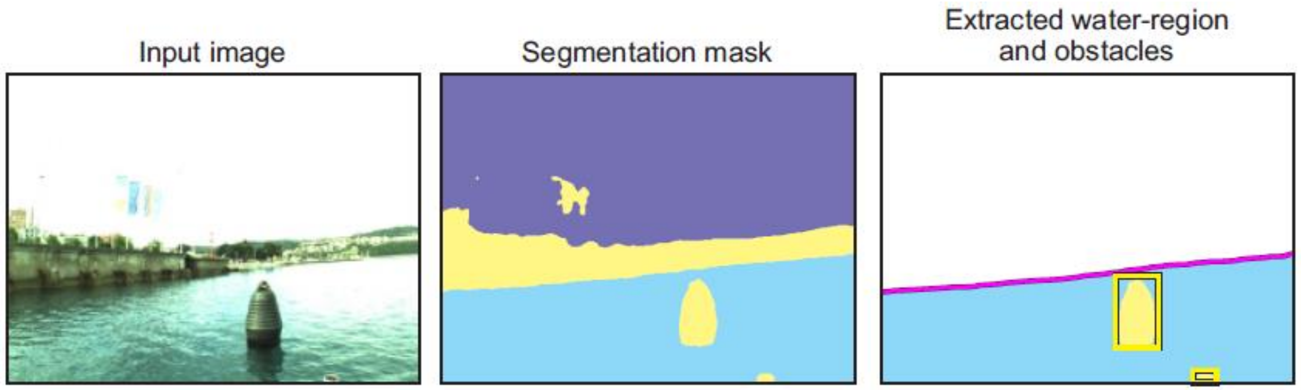


WaSR architecture





Architecture	μ_{edg}	TP	FP	FN	F-measure
PSPNet [12]	13.8 (16.0)	5886	4359	431	71.1
SegNet [35]	13.5 (18.5)	5834	2139	483	81.7
DL2$_{\text{NOCRF}}$ [11]	12.8 (21.4)	3946	**227**	2371	75.2
DL3+ [14]	14.1 (20.9)	5311	2935	1006	72.9
BiSeNet [13]	12.4 (19.2)	5699	1894	618	81.9
WaSR	**9.6** (18.5)	**6166**	679	**151**	**93.7**



WaSR results

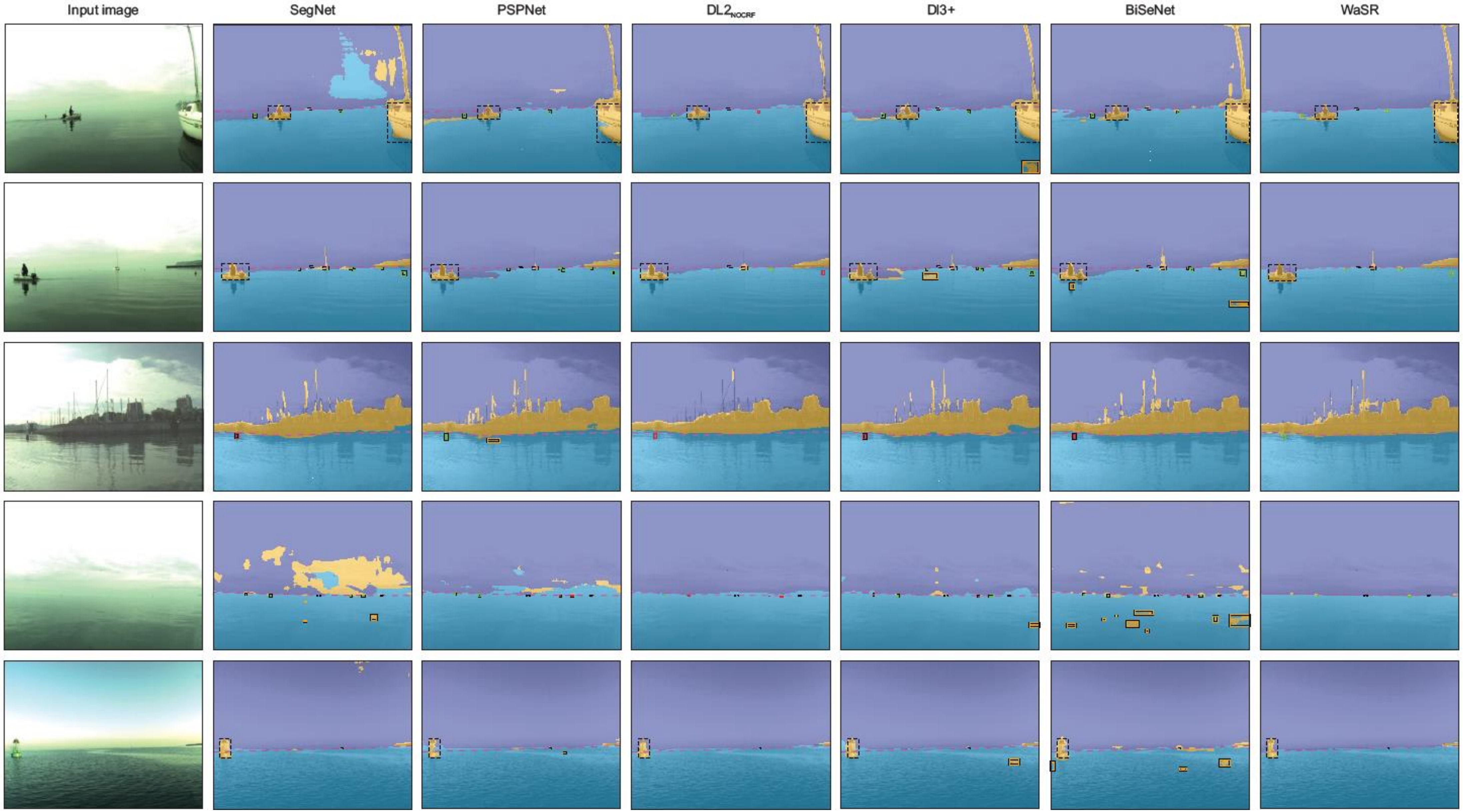
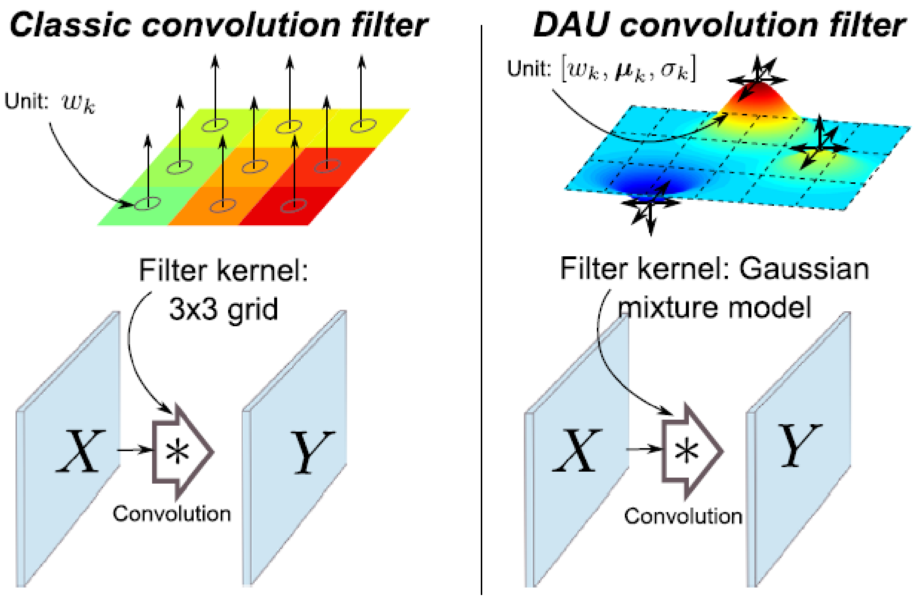


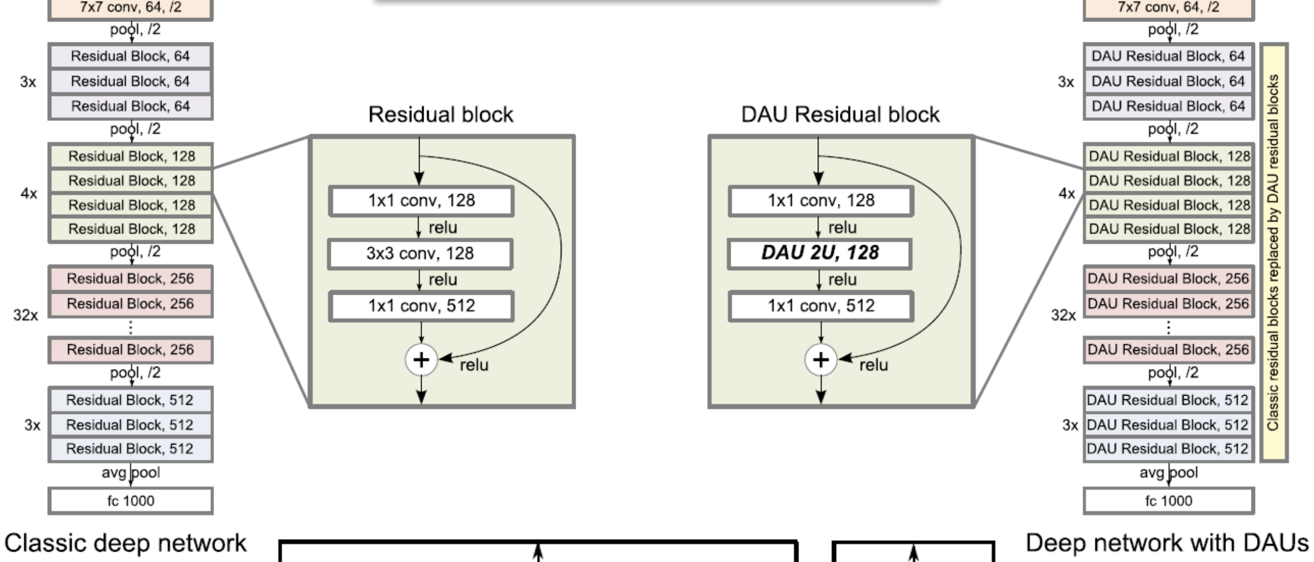
Image enhancement

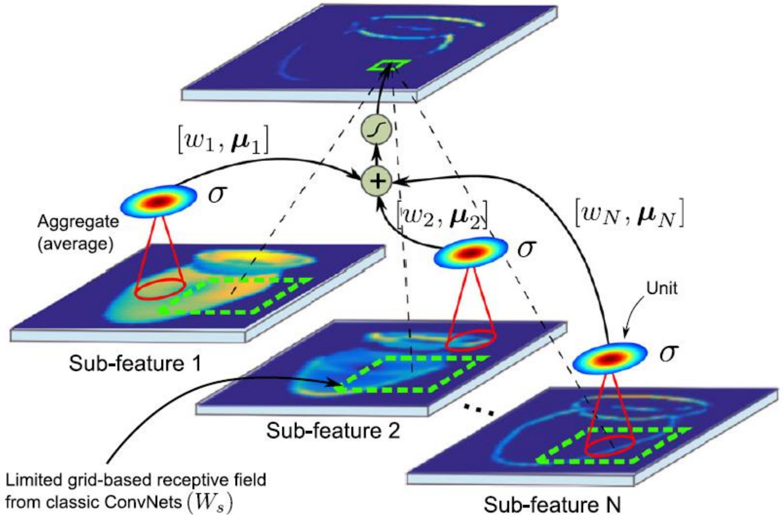
- Deblurring, super-resolution

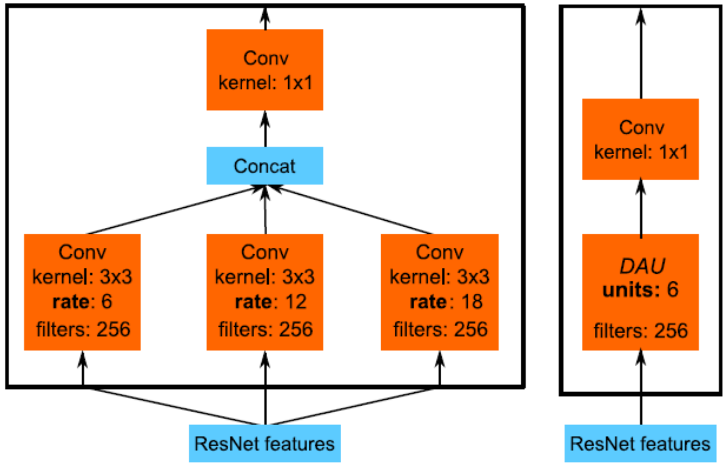
Spatially-Adaptive Filter Units



Tabernik et. al, 2020





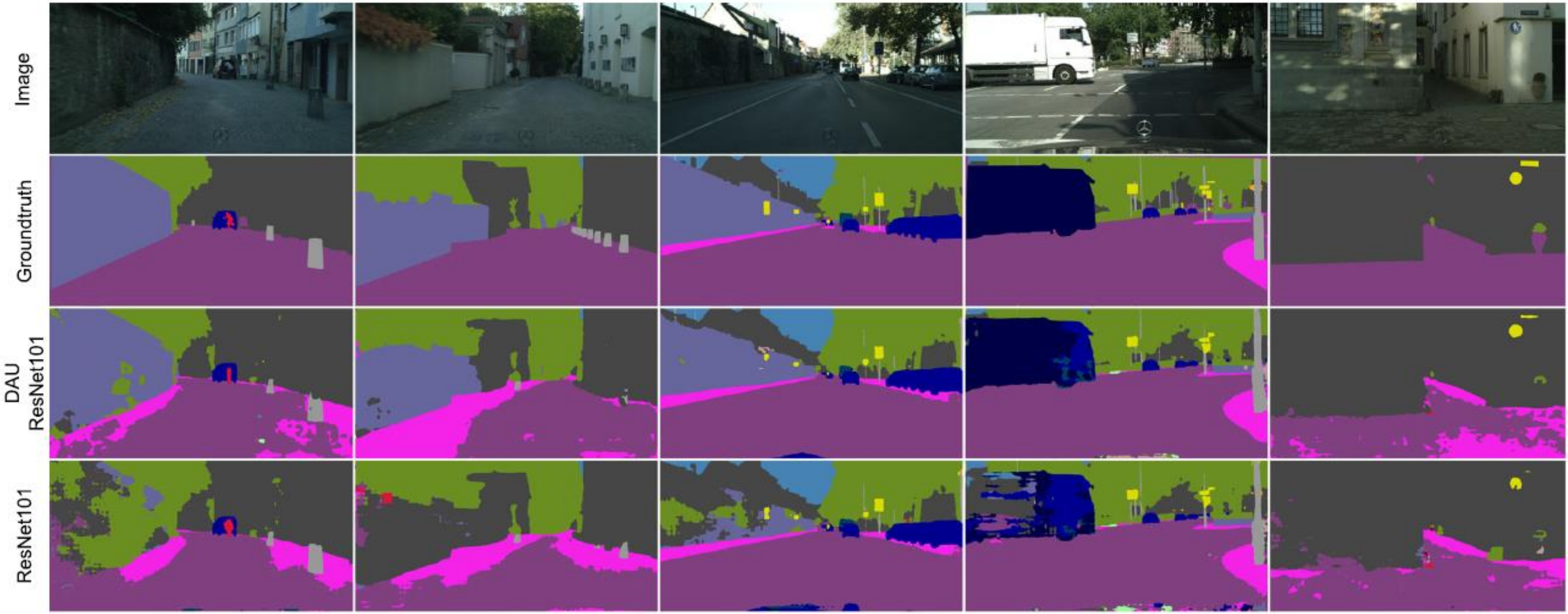


(a) Atrous Spatial Pyramid Pooling pathways **(b)** DAUs pathway

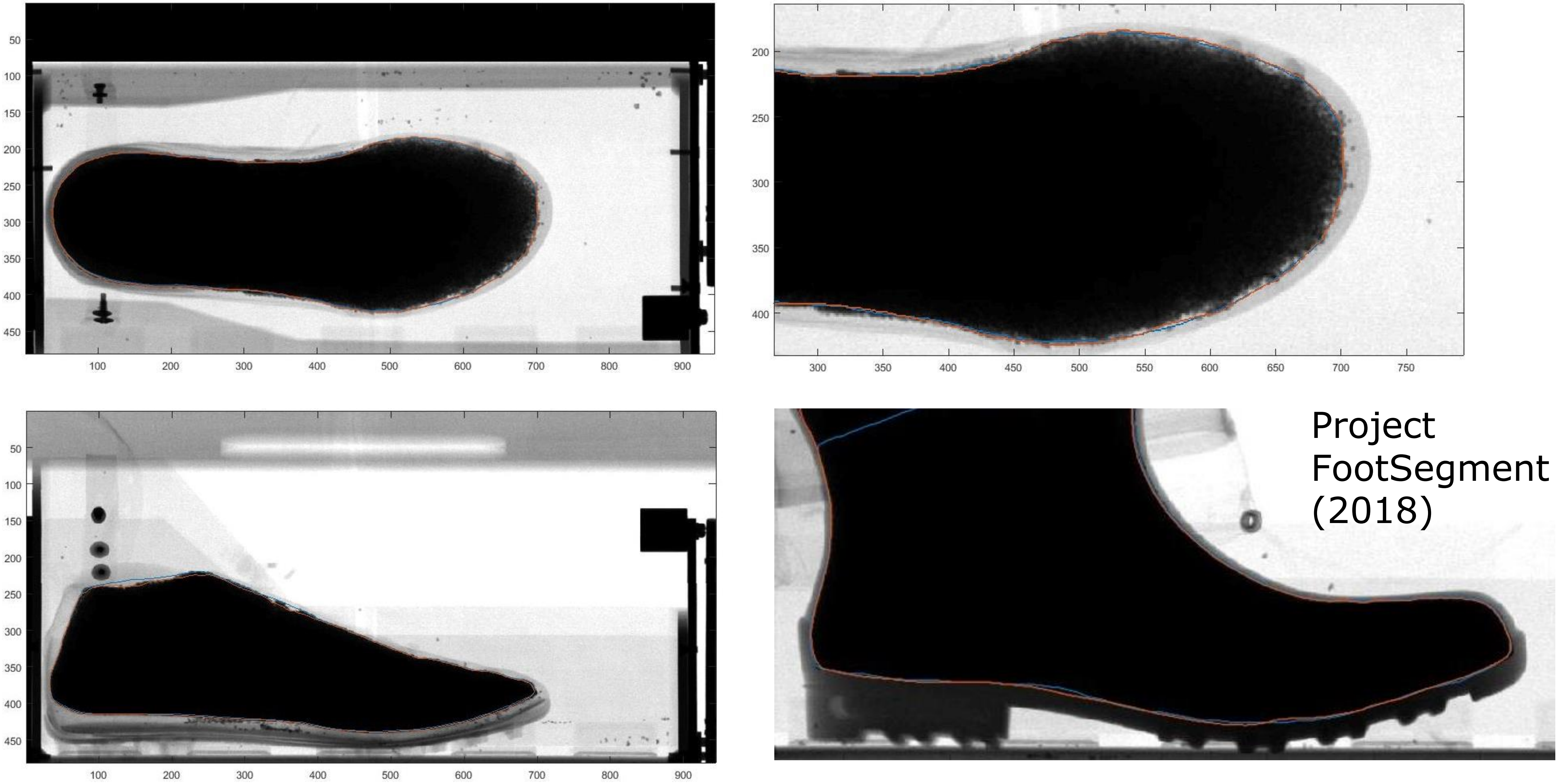

Original
Original
DAU-SNR-Deblur (our)
DAU-SNR-Deblur (our)


Original
Original
DAU-SNR-Deblur (our)
DAU-SNR-Deblur (our)

Semantic segmentation with DAUs

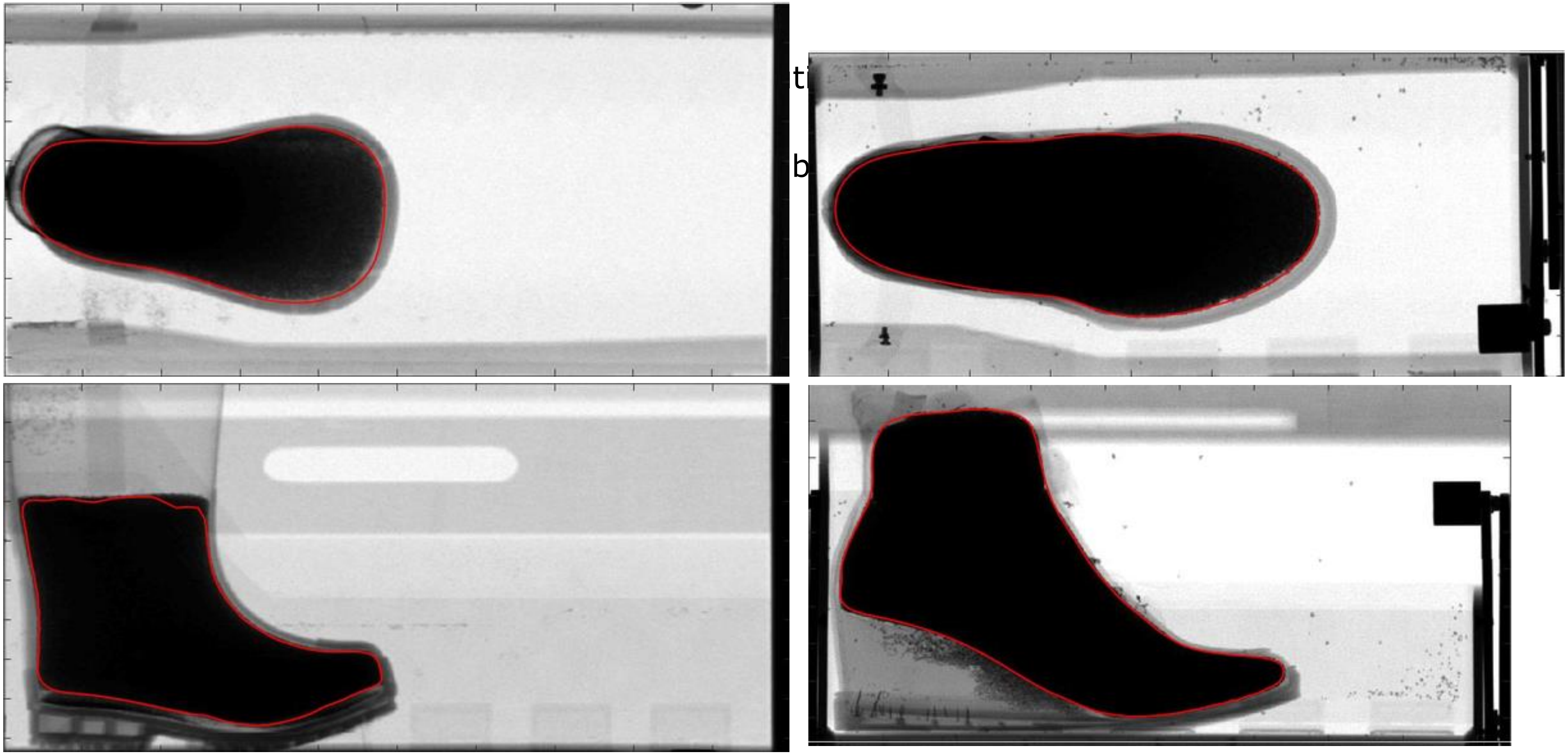


Segmentation for semantic edge detection



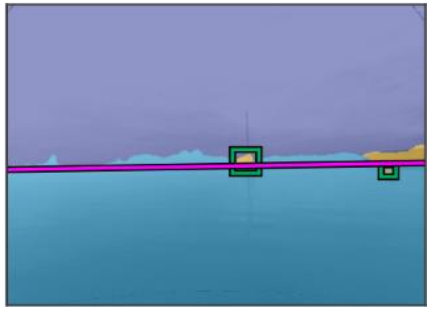
Project FootSegment (2018)

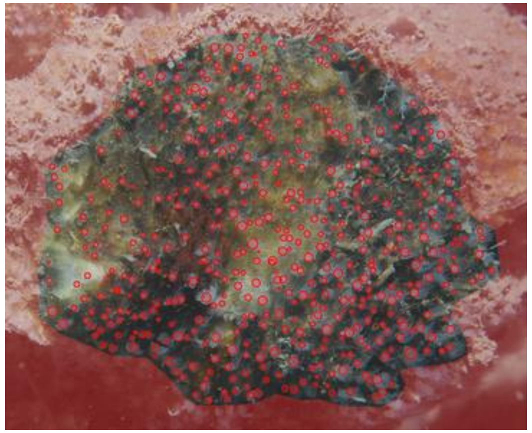
Segmentation for semantic edge detection

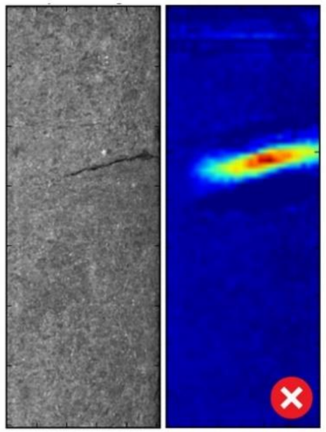


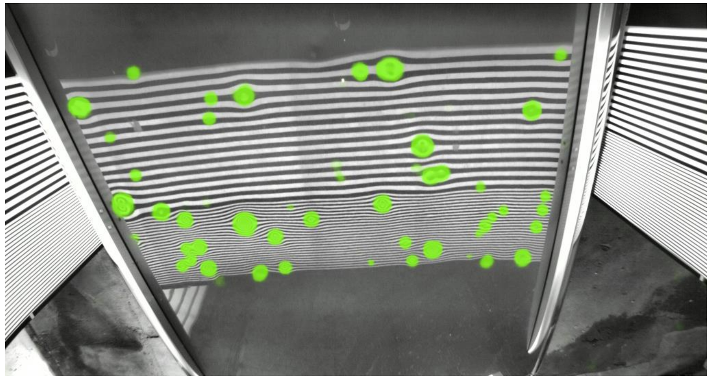
Segmentation for various computer vision tasks

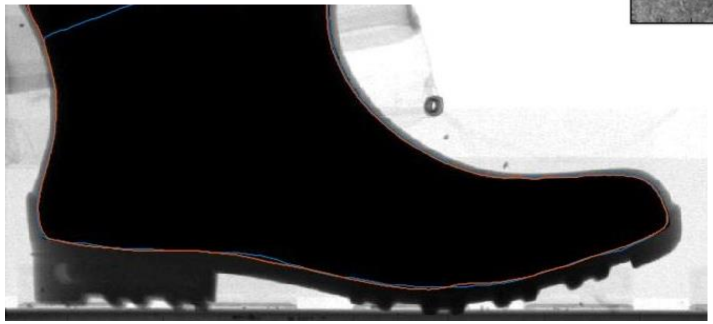
- Segmentation is very useful
  - For various applications
- In combination with classification and other problem-dependent loss functions
  - Elegant/general way of problem solving
- Data-driven learning-based problem solving
  - Key ingredient: training data!

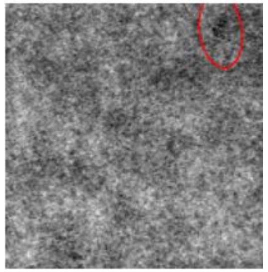




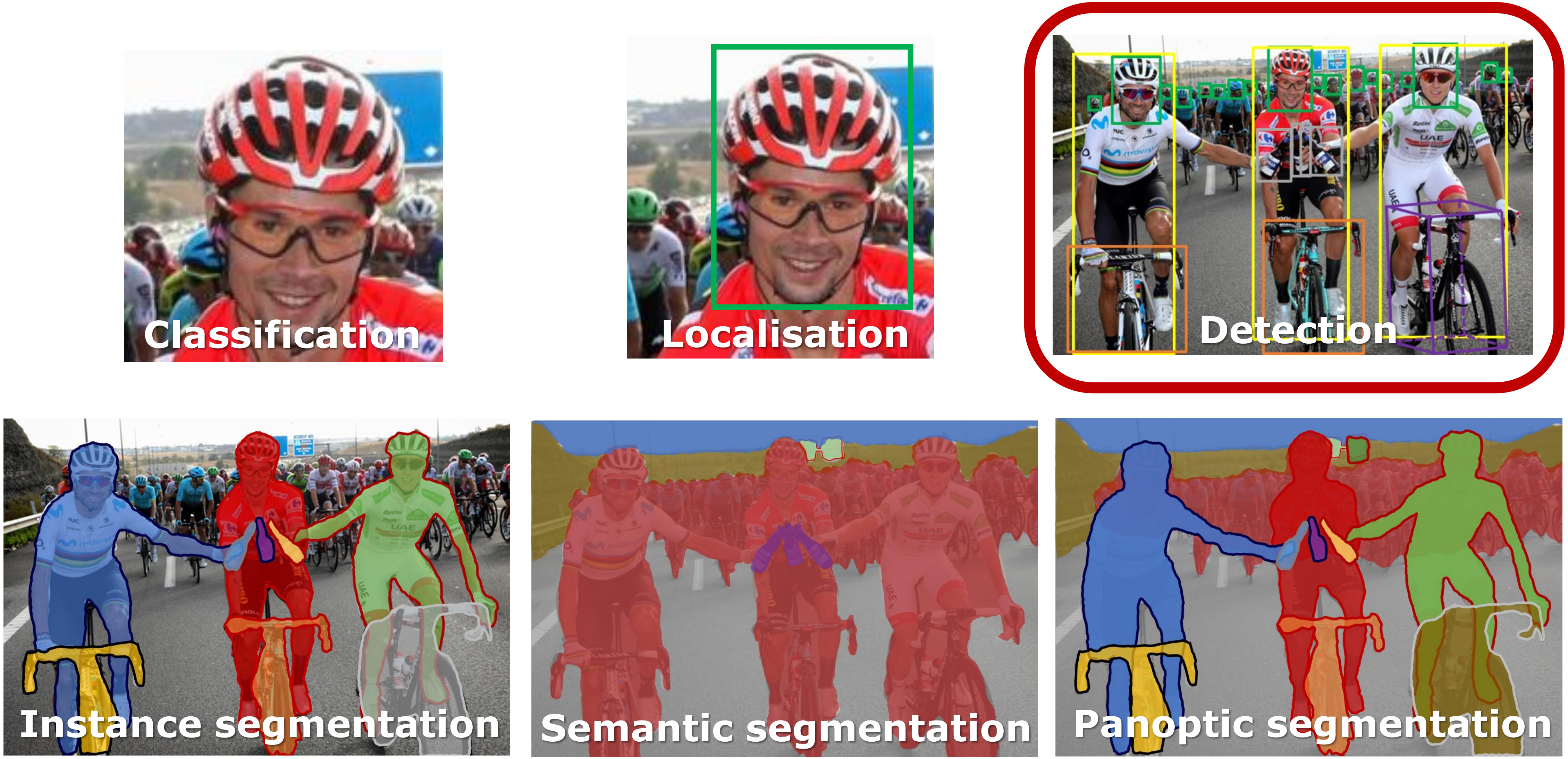






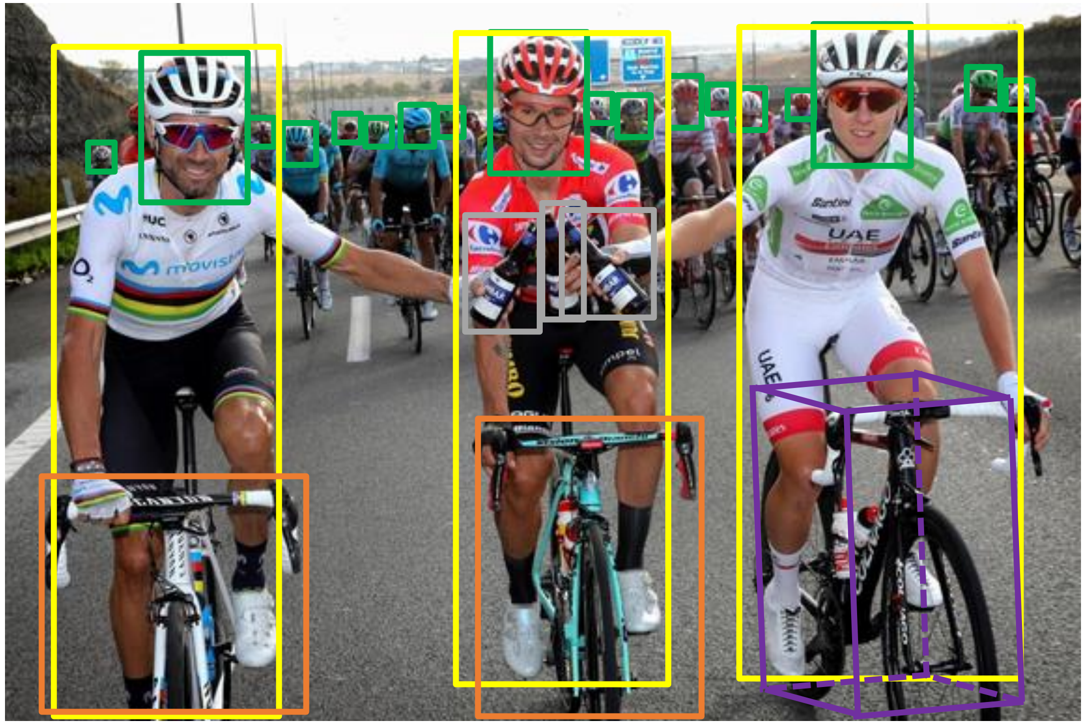


Detection


Classification
Localisation
Detection
Instance segmentation
Semantic segmentation
Panoptic segmentation

Detection

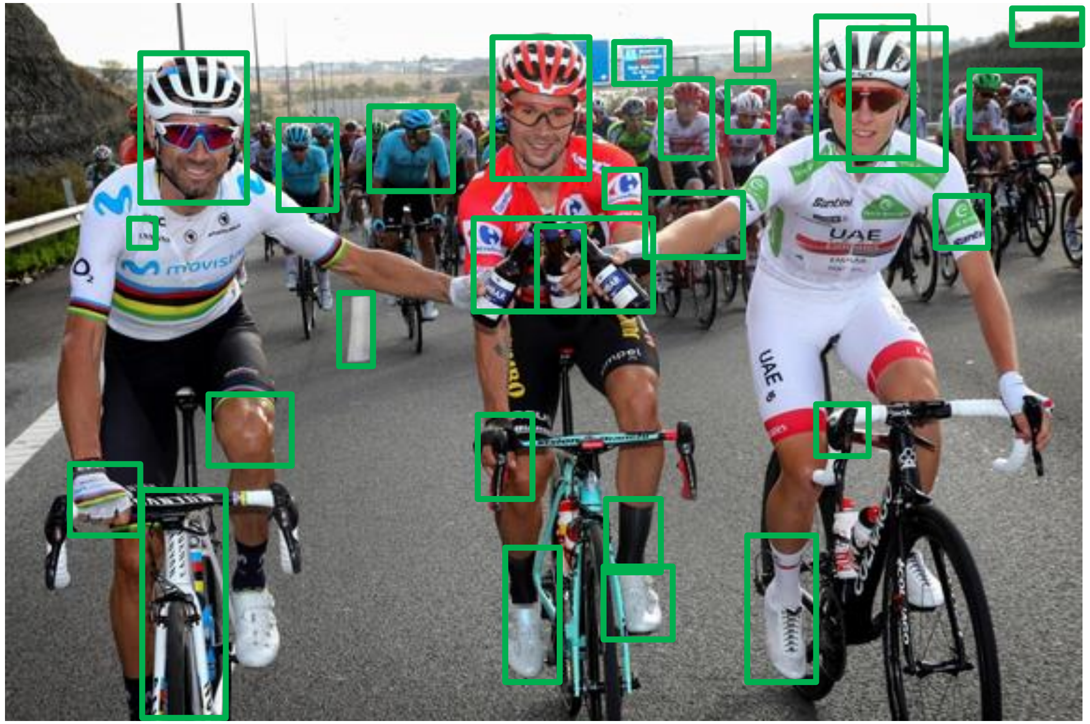
- Object detection – detect (localise and categorise) all the objects in the image
  - Unknown (arbitrary) number of objects
- Naive approach: Sliding window + classification
  - Too many locations, scales, aspect ratios!
  - Very expensive!

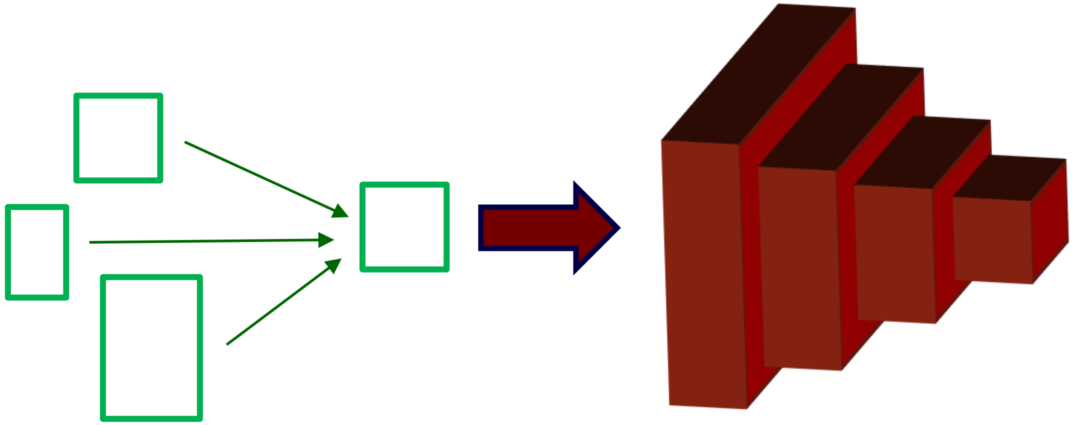


Region proposals

- Solution in early approaches:
  1. Find region proposals (regions of interest, potencial object candidates) – very fast
  2. Use CNN to classify these regions only (resize them to a predetermined size)

1. Many region proposals algorithms: objectness, selective search, BING, Edge boxes, etc.





Alexe et al., 2012

Uijlings et al., 2013

Cheng et al., 2014

Zitnick & Dollar, 2014

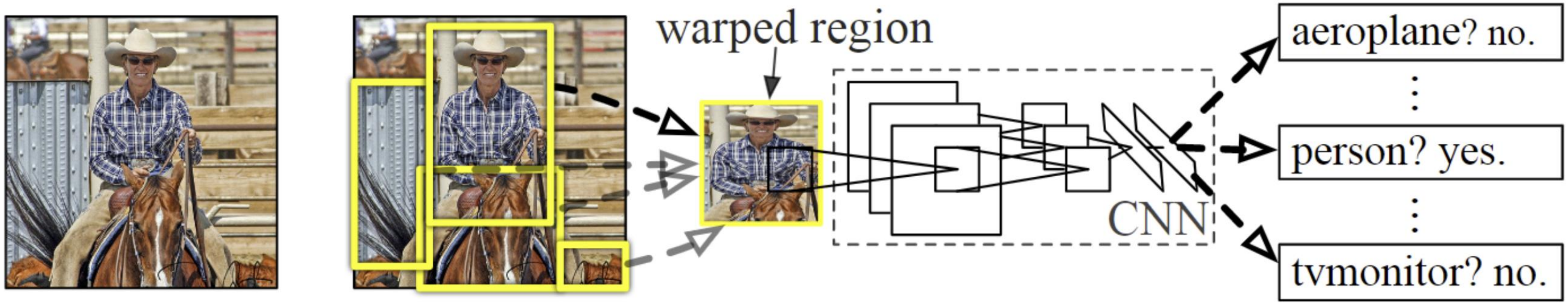
E.g. Overfeat

Sermanet et al., 2013

R-CNN

- Regions with CNN features - Region-based CNN
- Rich feature hierarchies for accurate object detection and semantic segmentation
- CNN as feature extraction only (ImageNet pretrained)
  - Use external region proposals (Selective search)
  - Use external classifiers (on CNN features)
    - SVM classification
    - Bounding box regression
- SOTA in 2014
- Extremely slow!
  - Each region passed through CNN

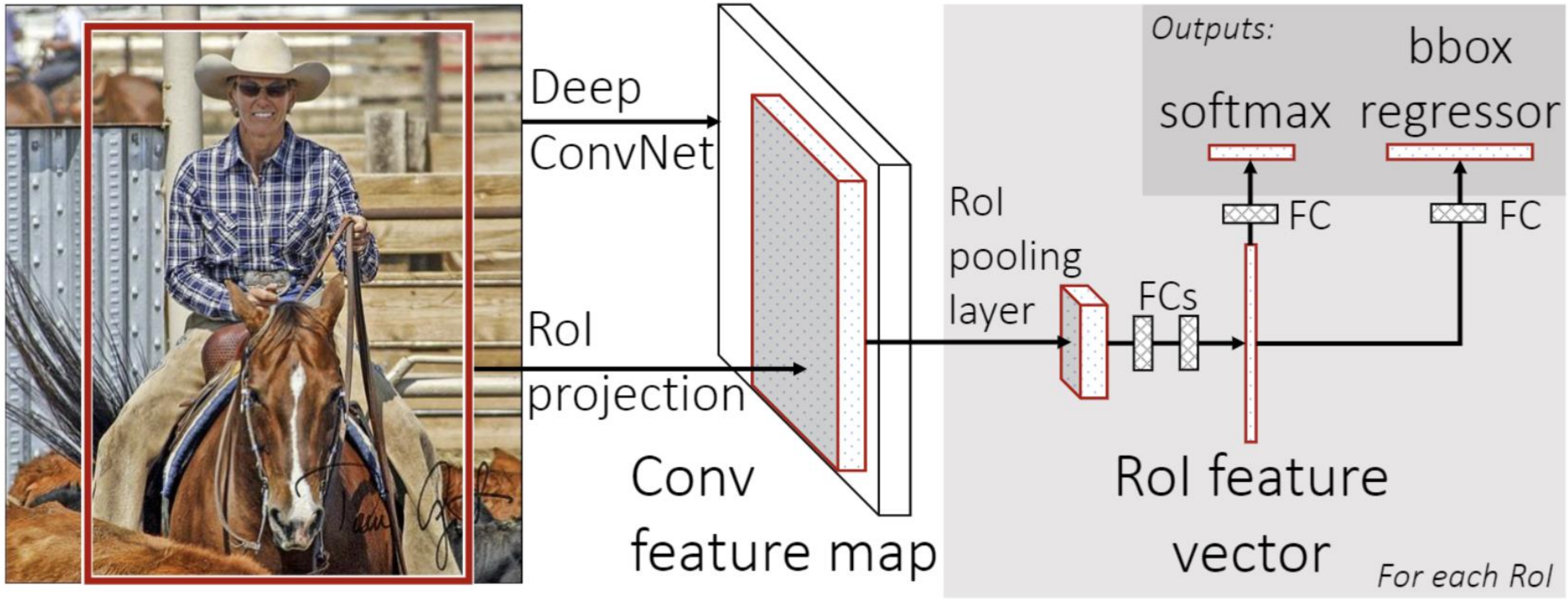
Girshick et al., 2014



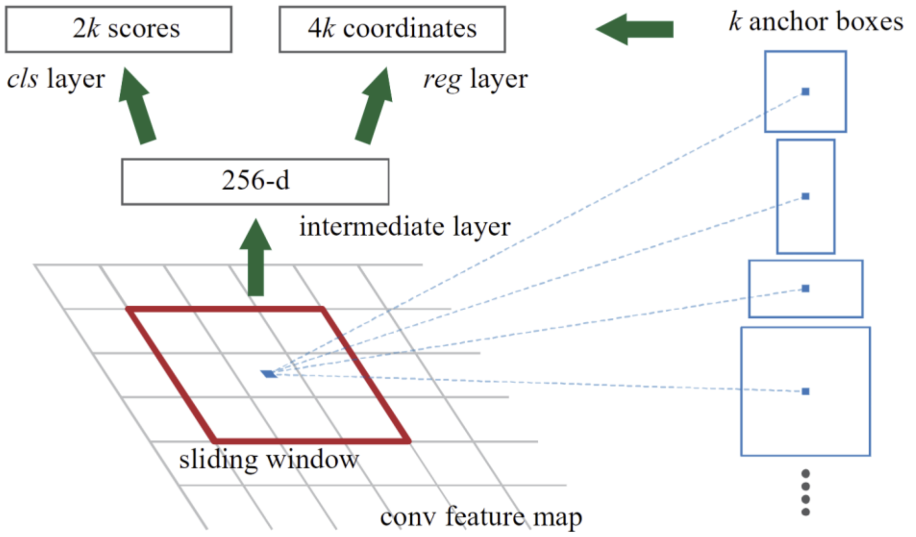
Fast R-CNN

- Fast Region-based Convolutional Network
- Still external region proposals
- Detection on CNN features
  - Images passed through CNN only once
  - RoI pooling – project RoIs to CNN features
    - Snap to grid + maxPooling
- Faster than R-CNN, however still slow
  - Due to external region proposal method
- SOTA in 2015

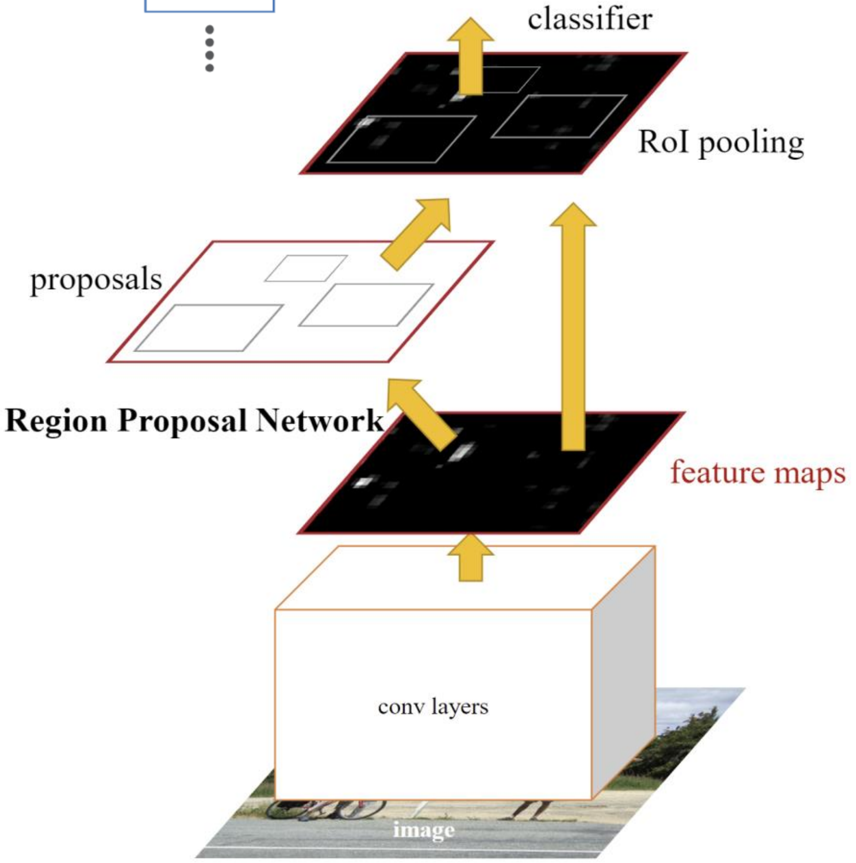
Girshick, 2015



Faster R-CNN

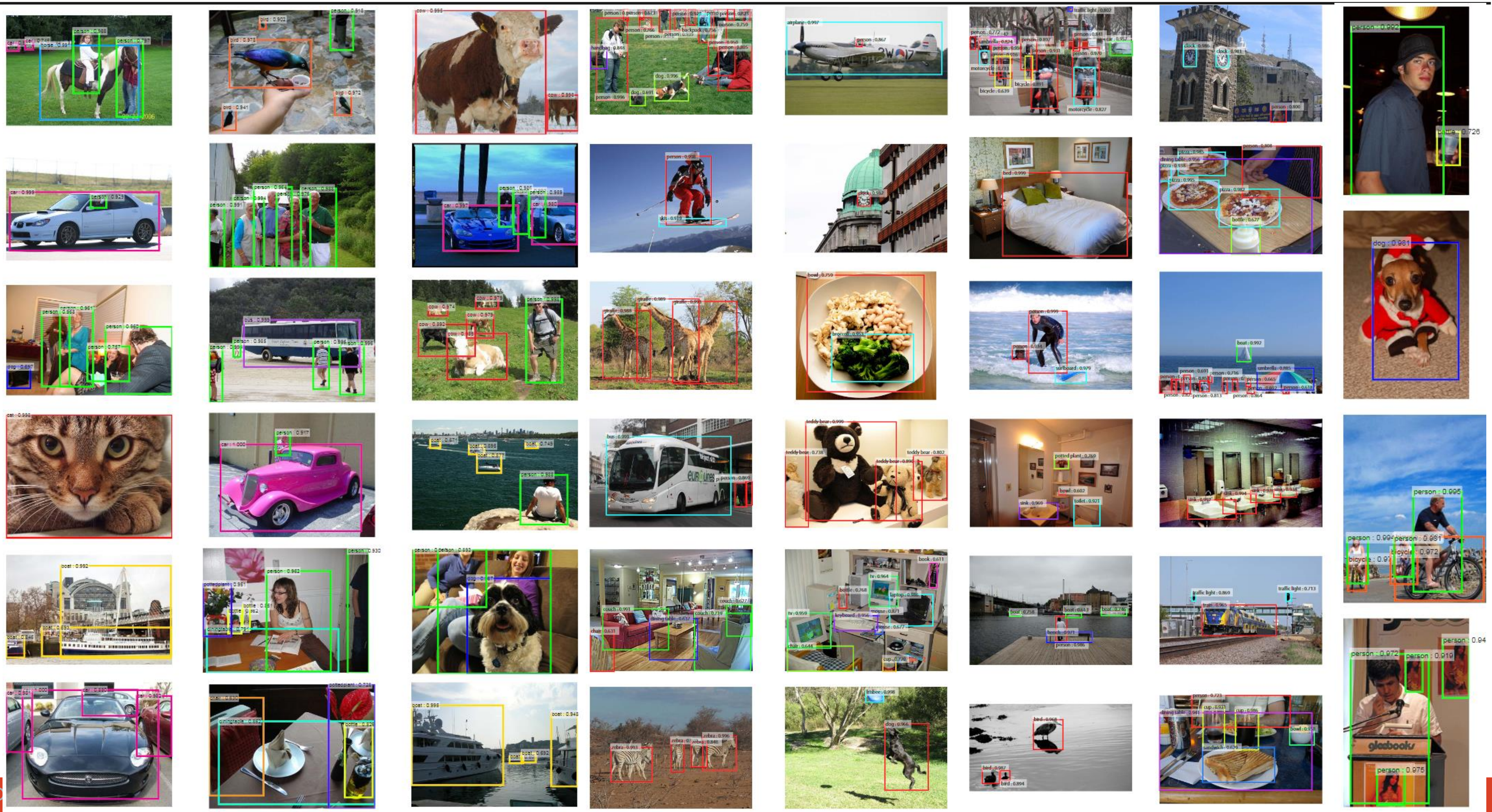


- Region Proposal Network
  - Included in the method
  - Anchor boxes
  - Sliding window on feature map
- Two stage method (four losses)
  - Detect region proposals
    - Object bounds - RP cls loss (is object?)
    - Objectness score - RP BB loss (bb corrections)
  - Classify individual proposals
    - Cls loss (what it is?)
    - BB loss (refine RP BB)
- Alternating / end-to-end learning
- Significantly faster than Fast R-CNN
- SOTA in 2015

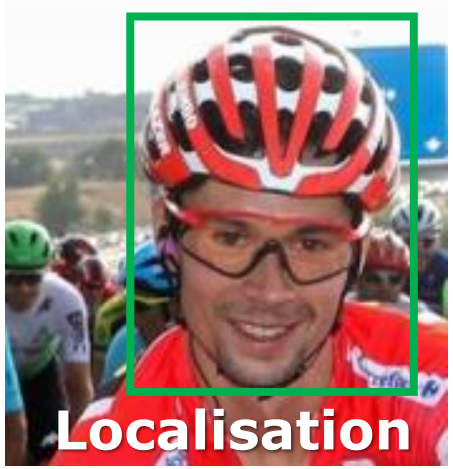


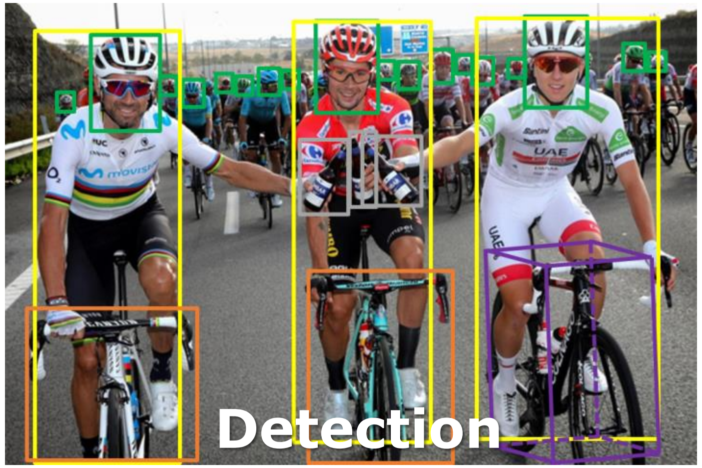
Ren et al., 2015

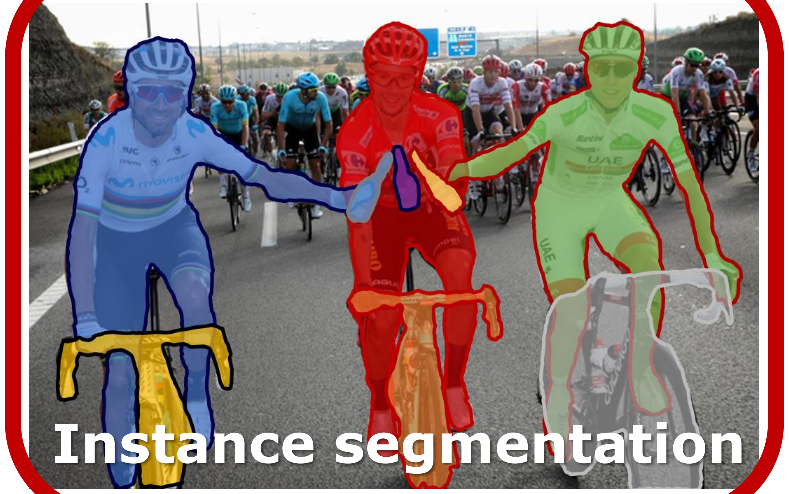
Faster R-CNN results

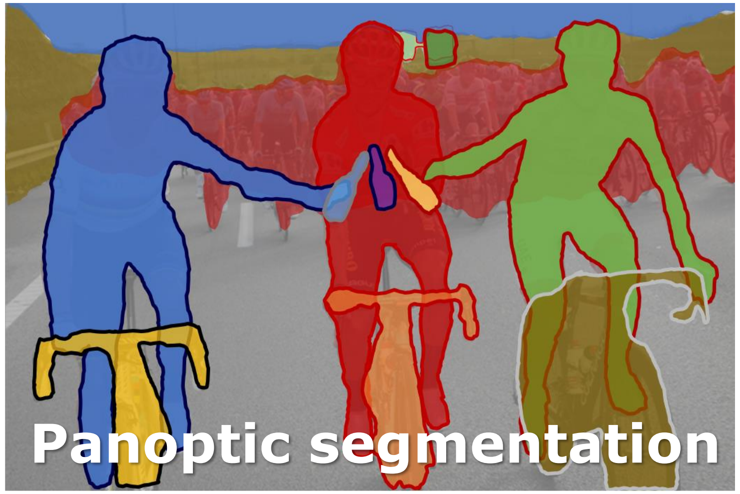


Instance segmentation



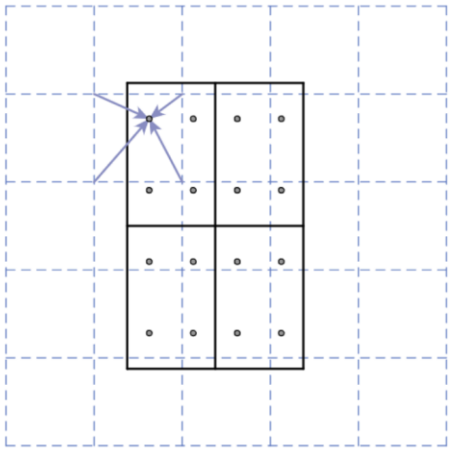


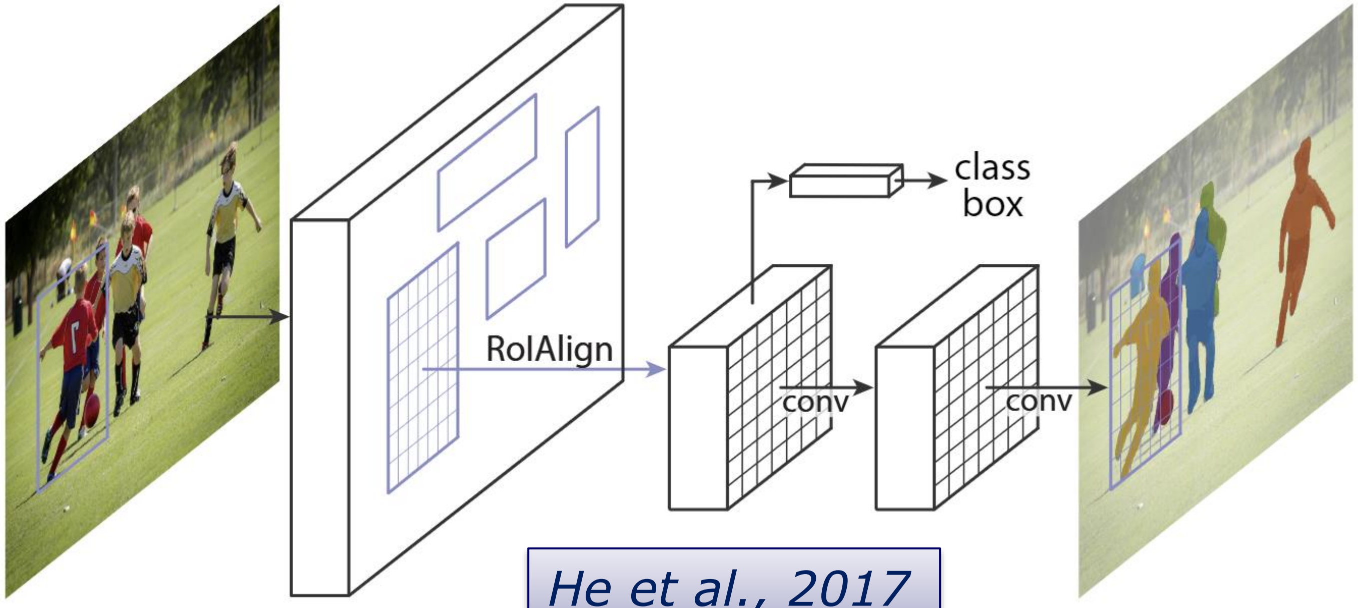




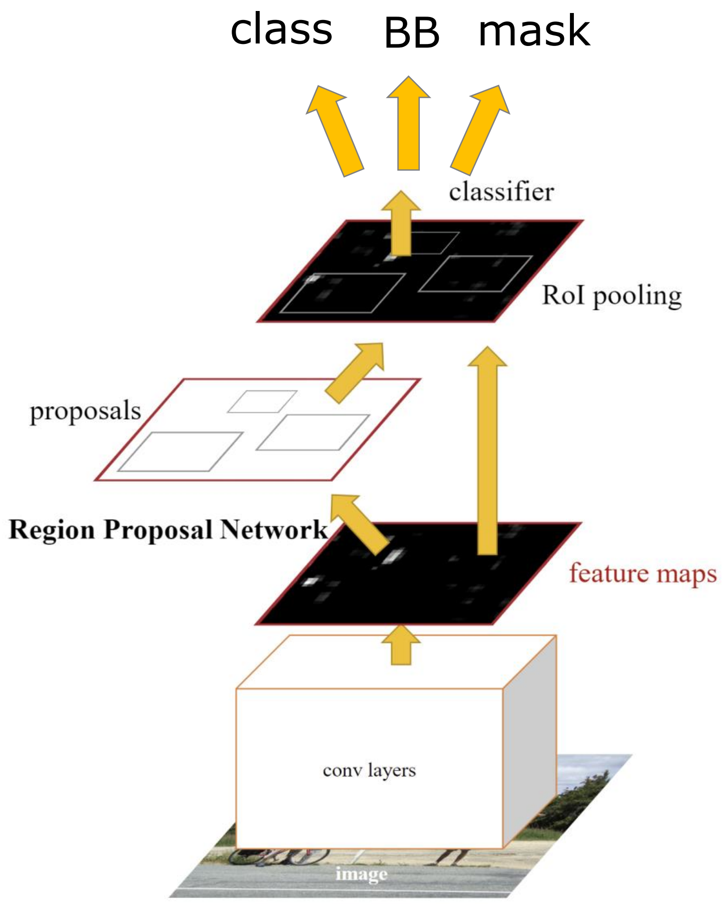
Mask R-CNN

- Add segmentation head
  - Additional segmentation loss
  - Produces segmentation mask for every RoI
- RoI align
- Other extensions possible





He et al., 2017



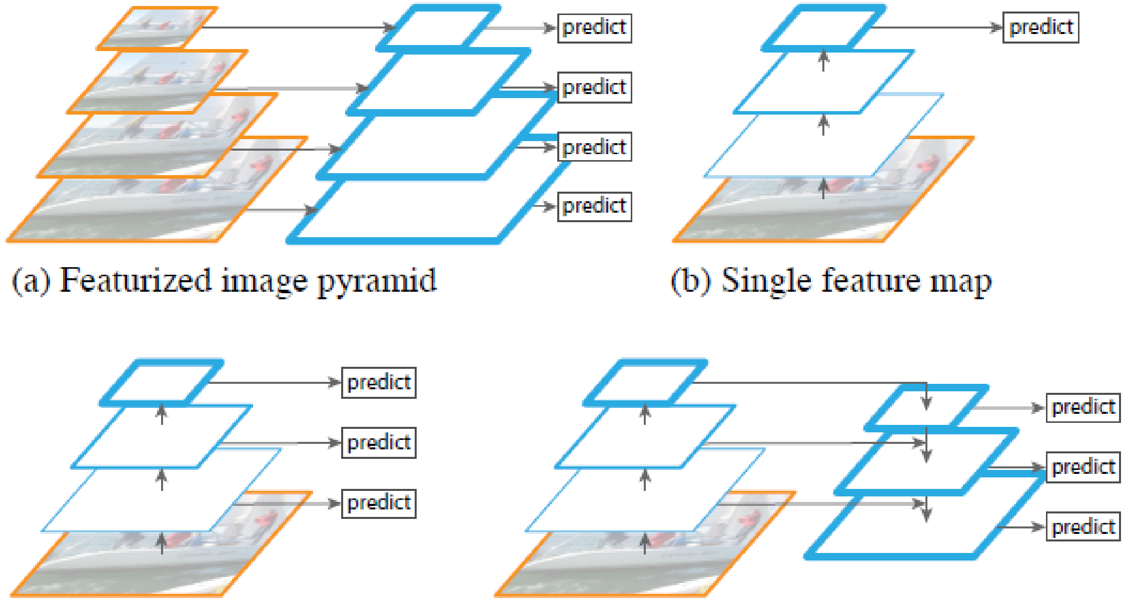
Mask-RCNN results

Mask R-CNN extensions

- Add task-specific heads
- E.g. human keypoint prediction
  - Key-point head
  - Predict 17 masks for the individual body parts

He et al., 2017

FPN - Feature Pyramid Network

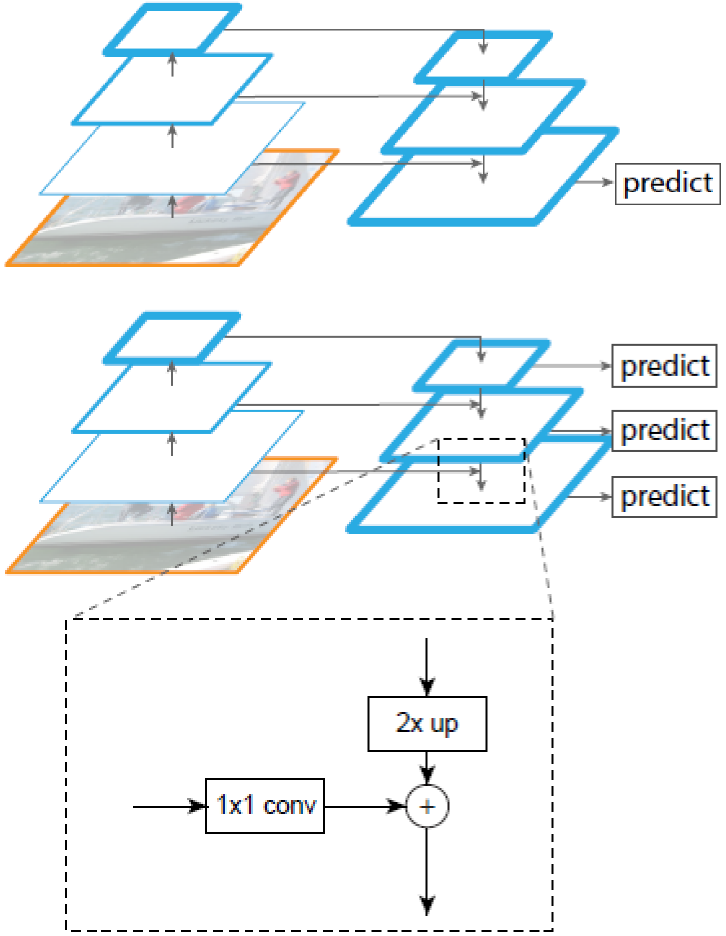


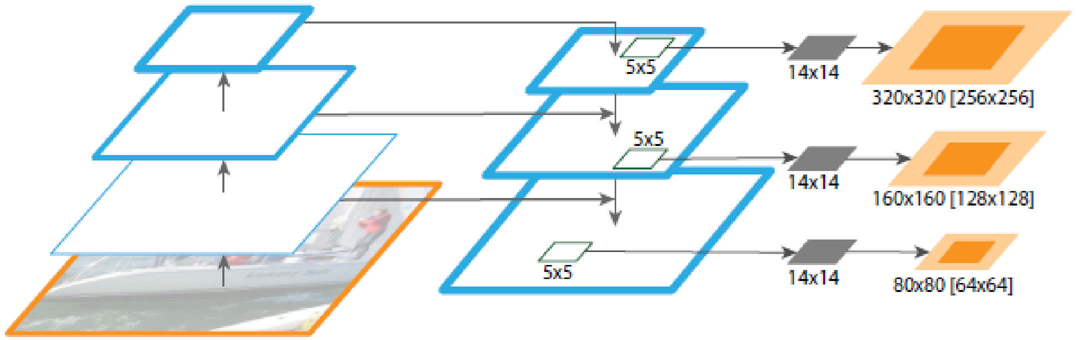
(a) Featurized image pyramid

(b) Single feature map

(c) Pyramidal feature hierarchy

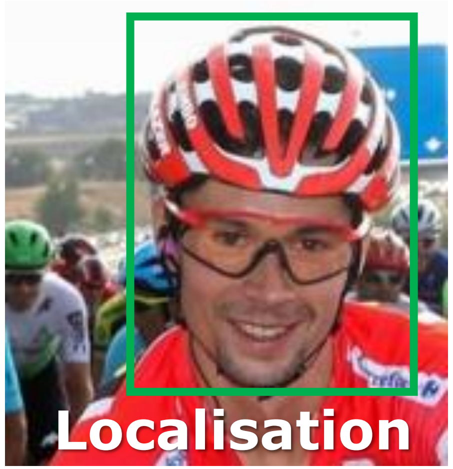
(d) Feature Pyramid Network

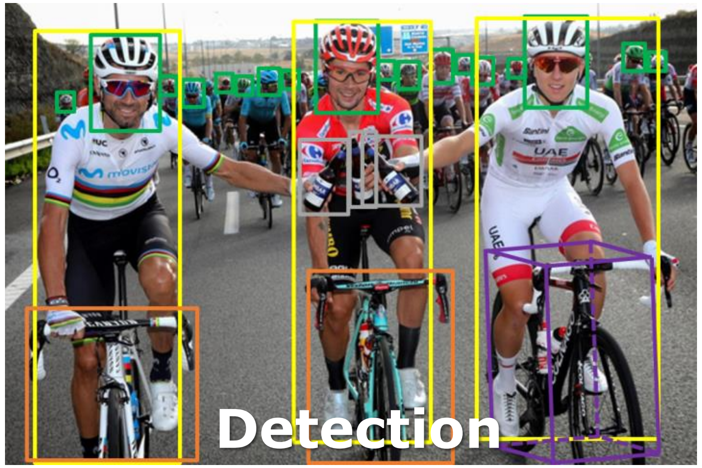


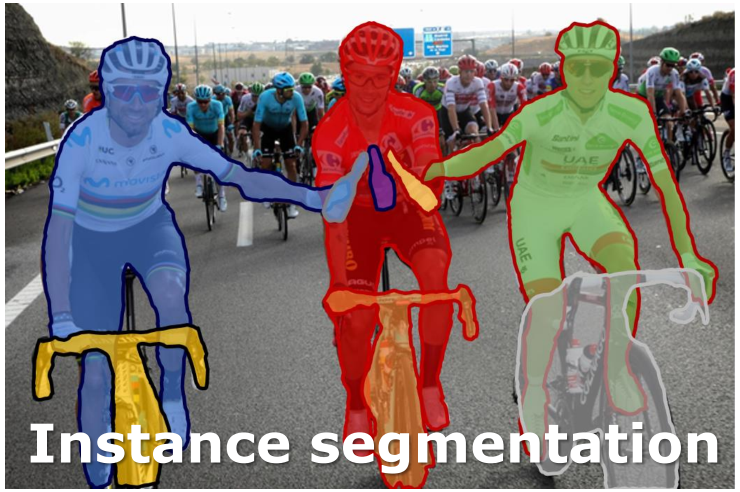


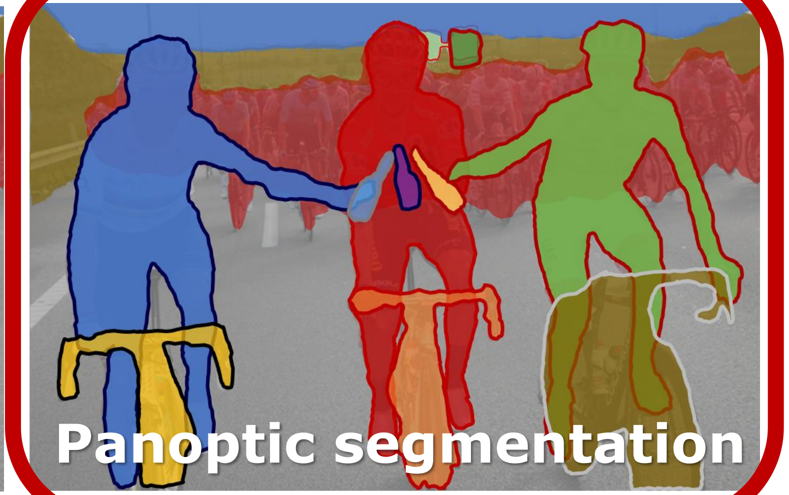
Lin et al., 2017

Panoptic segmentation



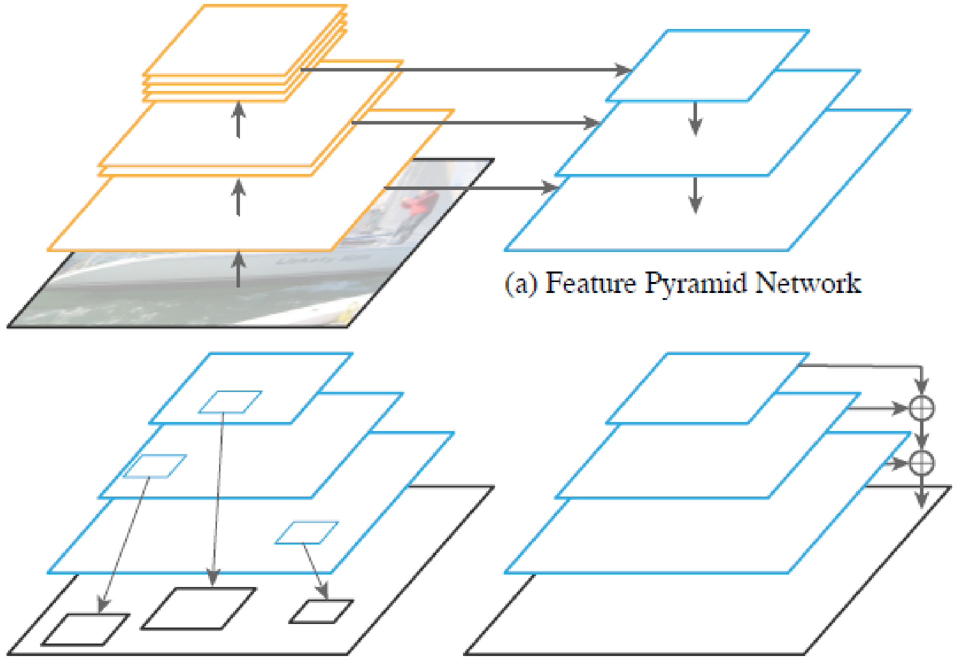


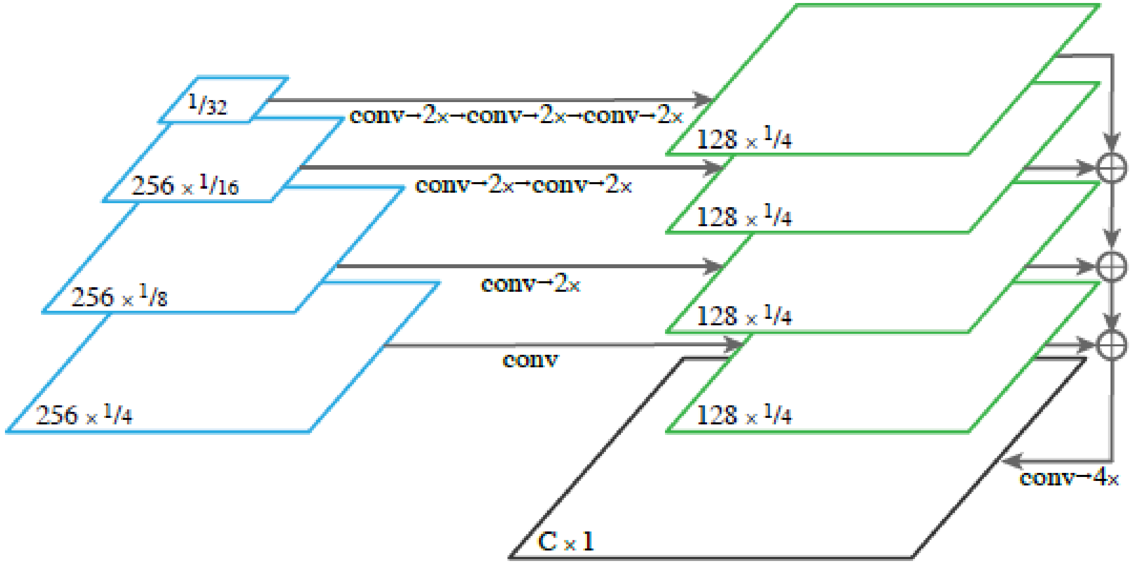




Panoptic Feature Pyramid Networks

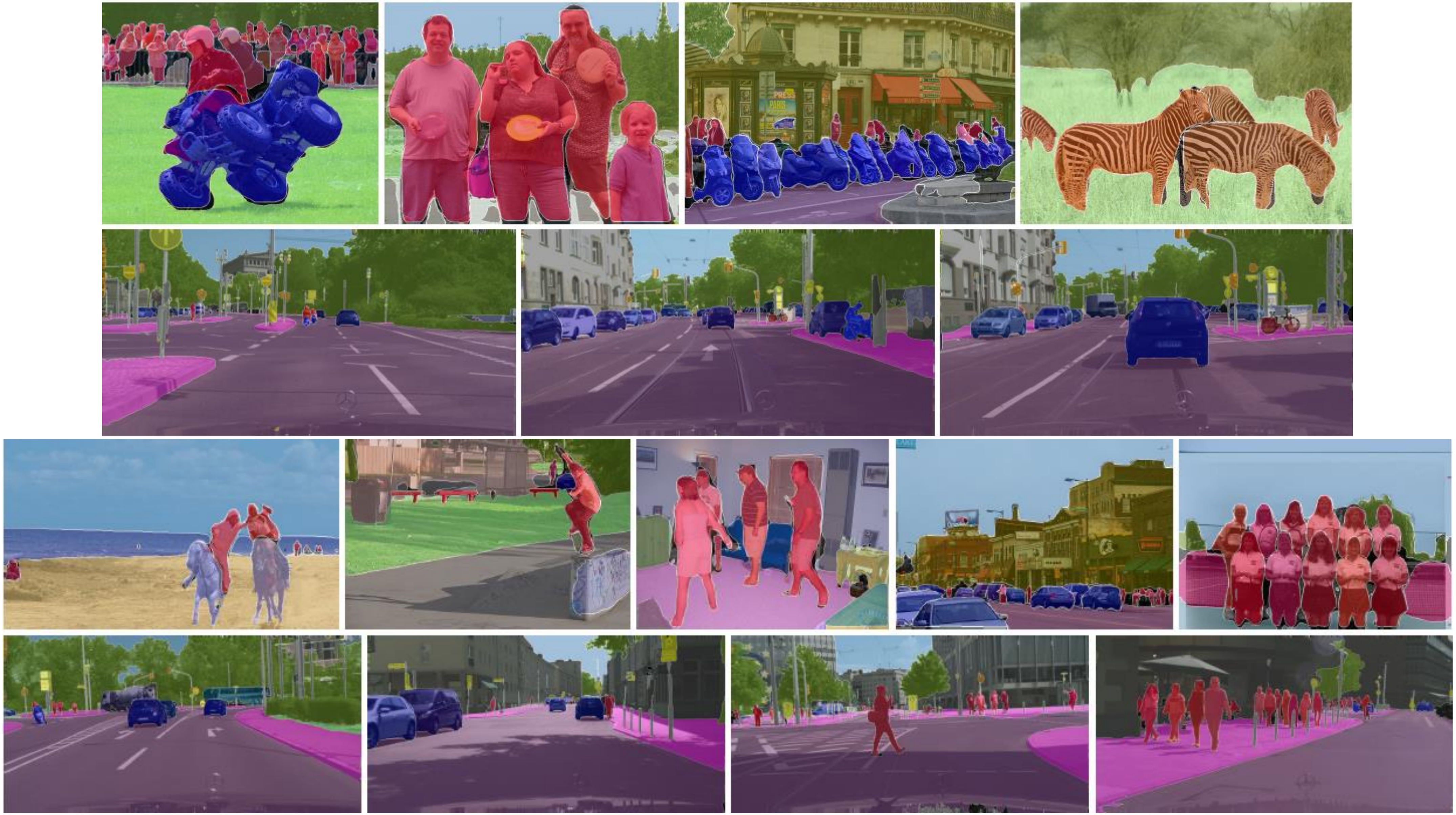
- Instance segmentation + semantic segmentation
- Mask-RCNN + FPN + semantic segmentation branch
- A single network





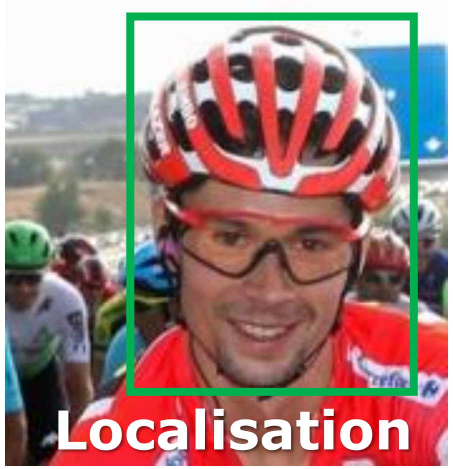
Kirilov et al., 2019

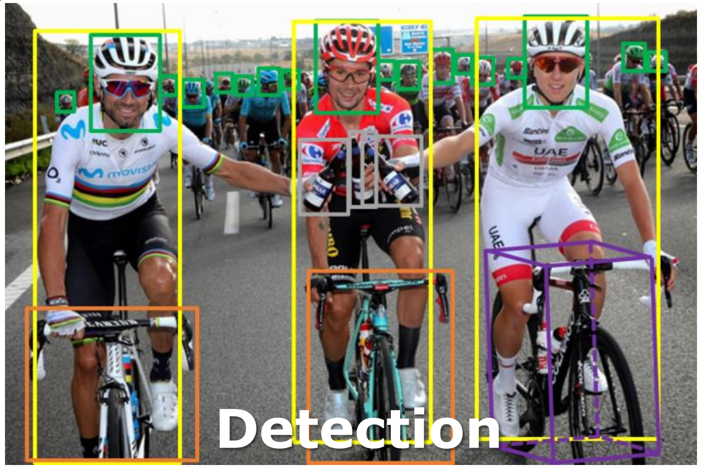
Panoptic FPN results

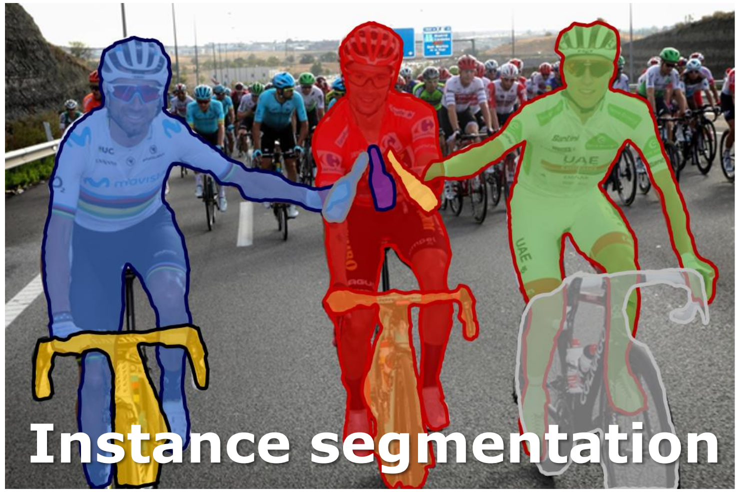


Detection

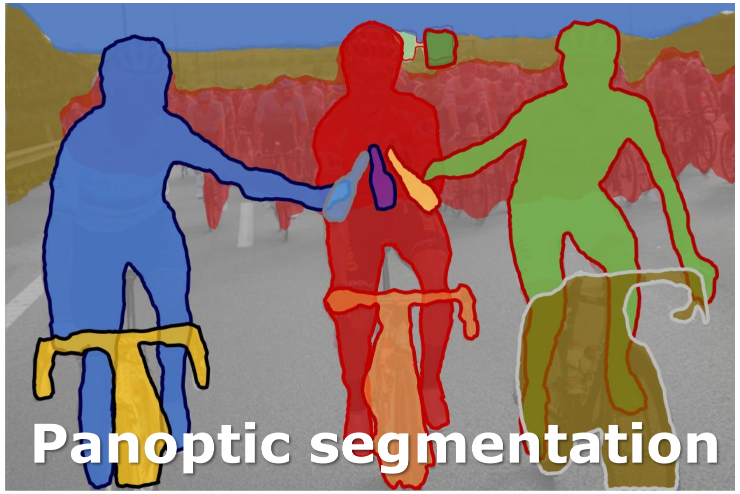

Classification


Localisation


Detection

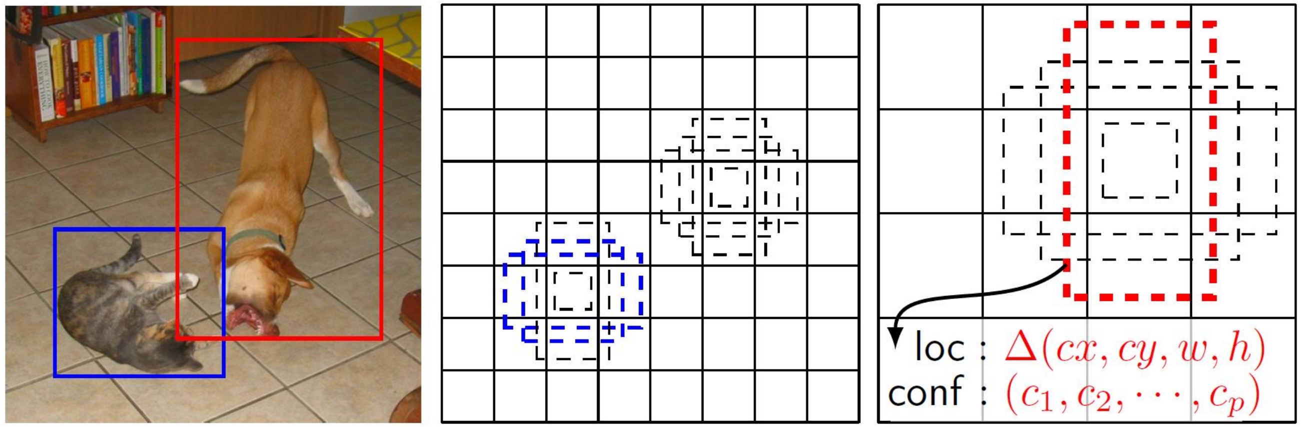

Instance segmentation

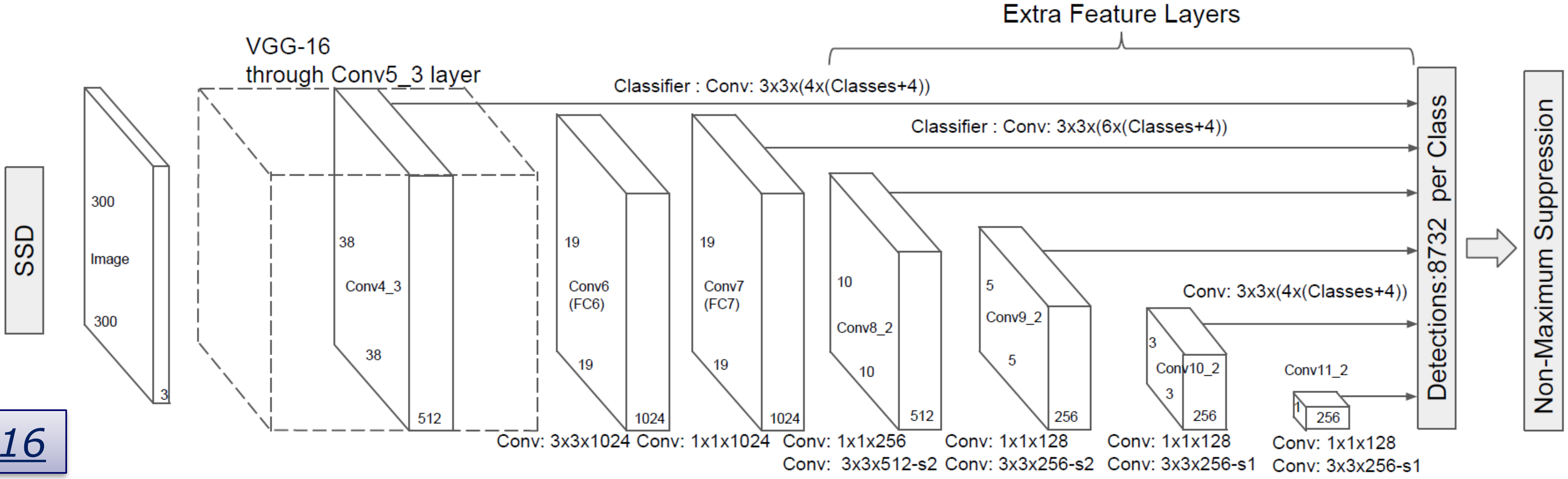

Semantic segmentation


Panoptic segmentation

SSD: Single Shot MultiBox Detector

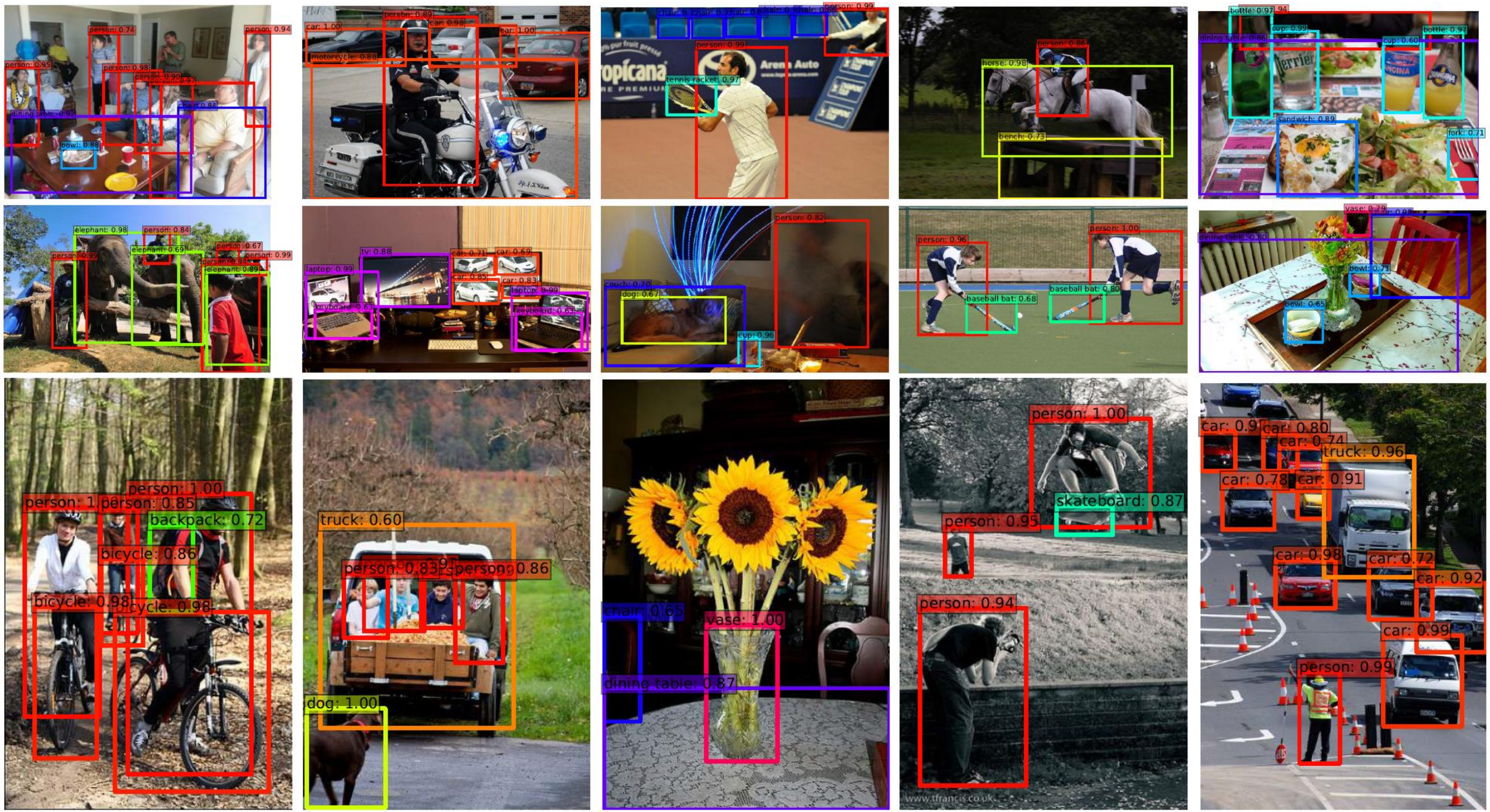
- Multi-scale feature maps for detection
- Convolutional predictors for detection
- Default boxes and aspect ratios
- Real time operation





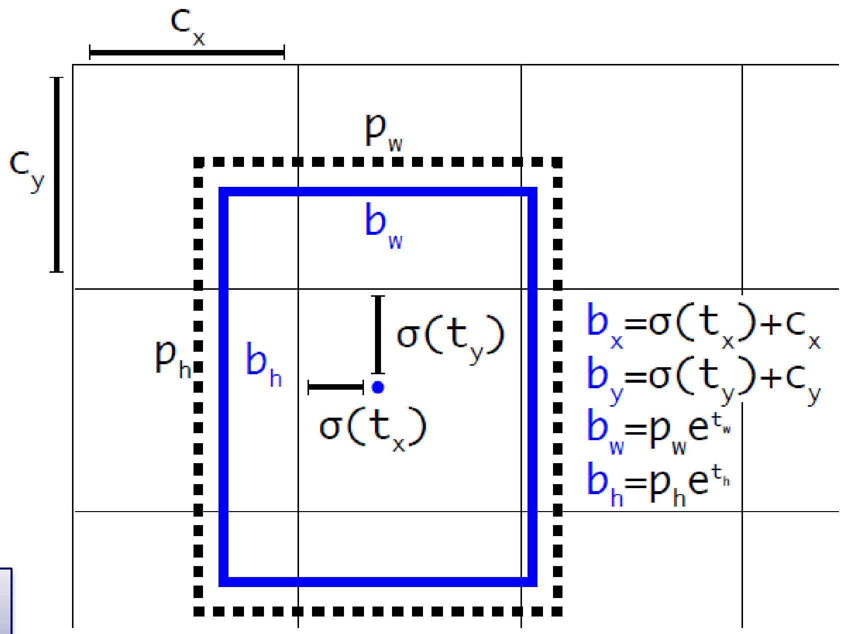
Liu et al., 2016

SSD results


person: 1.00
person: 0.85
backpack: 0.72
bicycle: 0.86
bicycle: 0.98
truck: 0.60
person: 0.83
person: 0.86
dog: 1.00
chair: 0.65
vase: 1.00
dining table: 0.87
person: 1.00
skateboard: 0.87
person: 0.95
person: 0.94
www.tfrancis.co.uk
car: 0.80
car: 0.74
truck: 0.96
car: 0.78
car: 0.91
car: 0.98
car: 0.72
car: 0.92
car: 0.99
person: 0.99

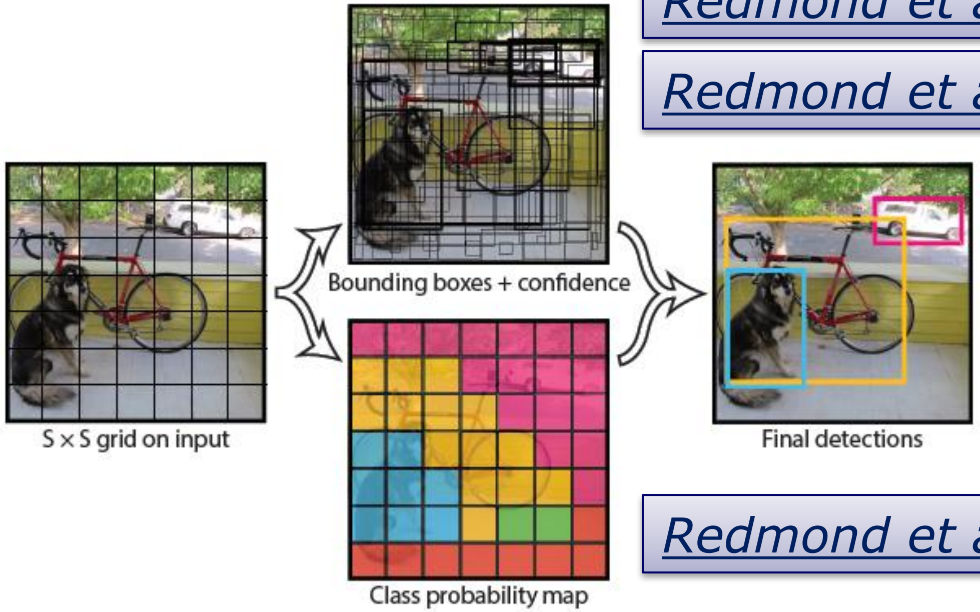
YOLOv3

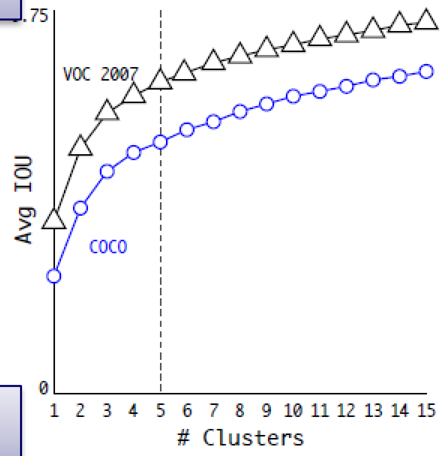
- You Only Look Once
- Prediction of bounding boxes on 3 scales
- 3 anchors as prior box shapes
- Prediction of objectness score for each BB
- Multilabel classification of each box
- Non-maxima suppression
- Real-time performance

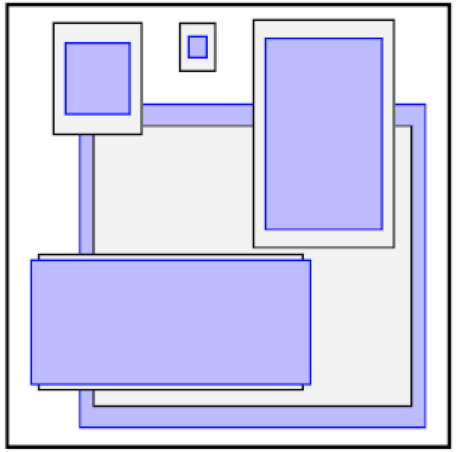


Redmond et al., 2016

Redmond et al., 2017

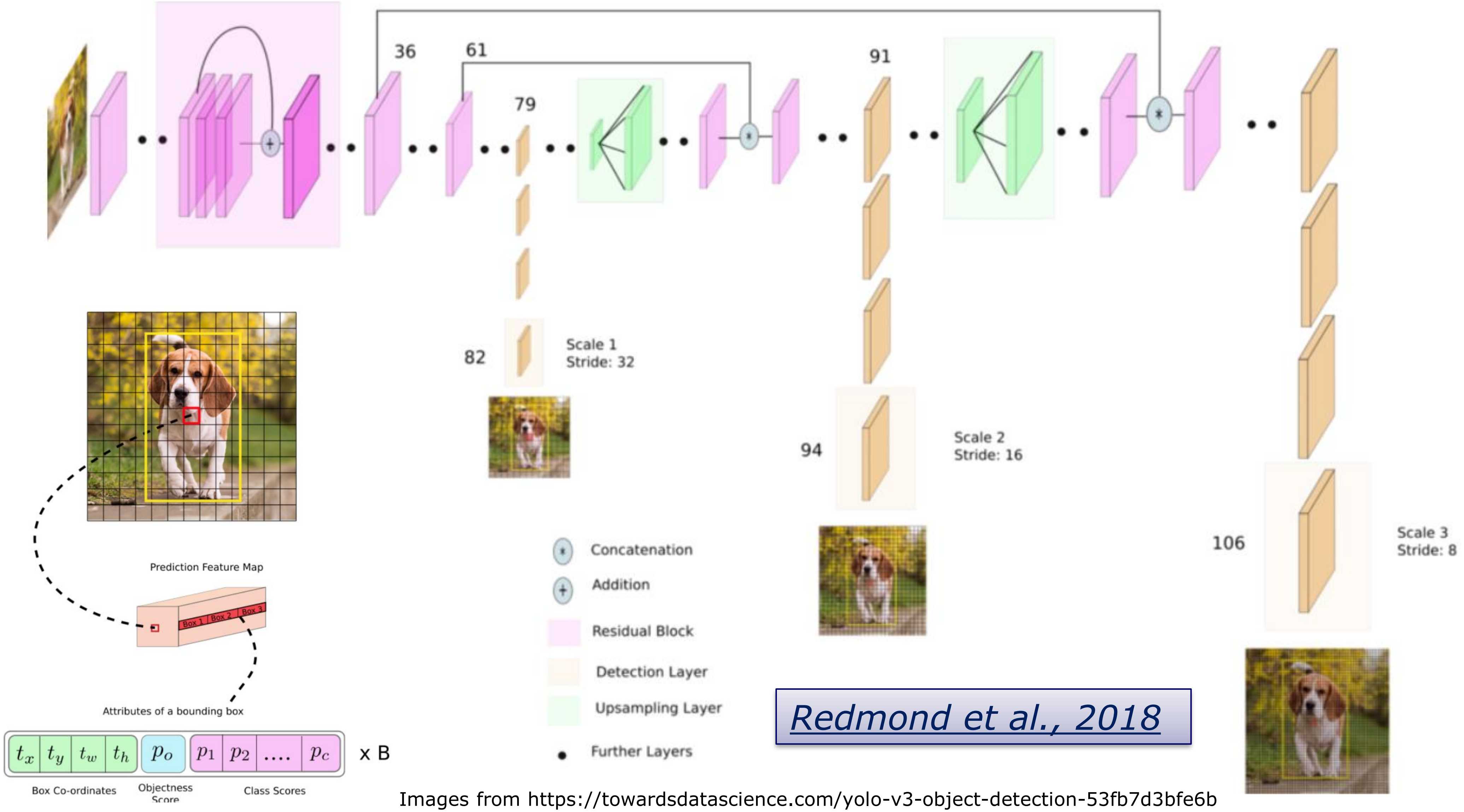






Redmond et al., 2018

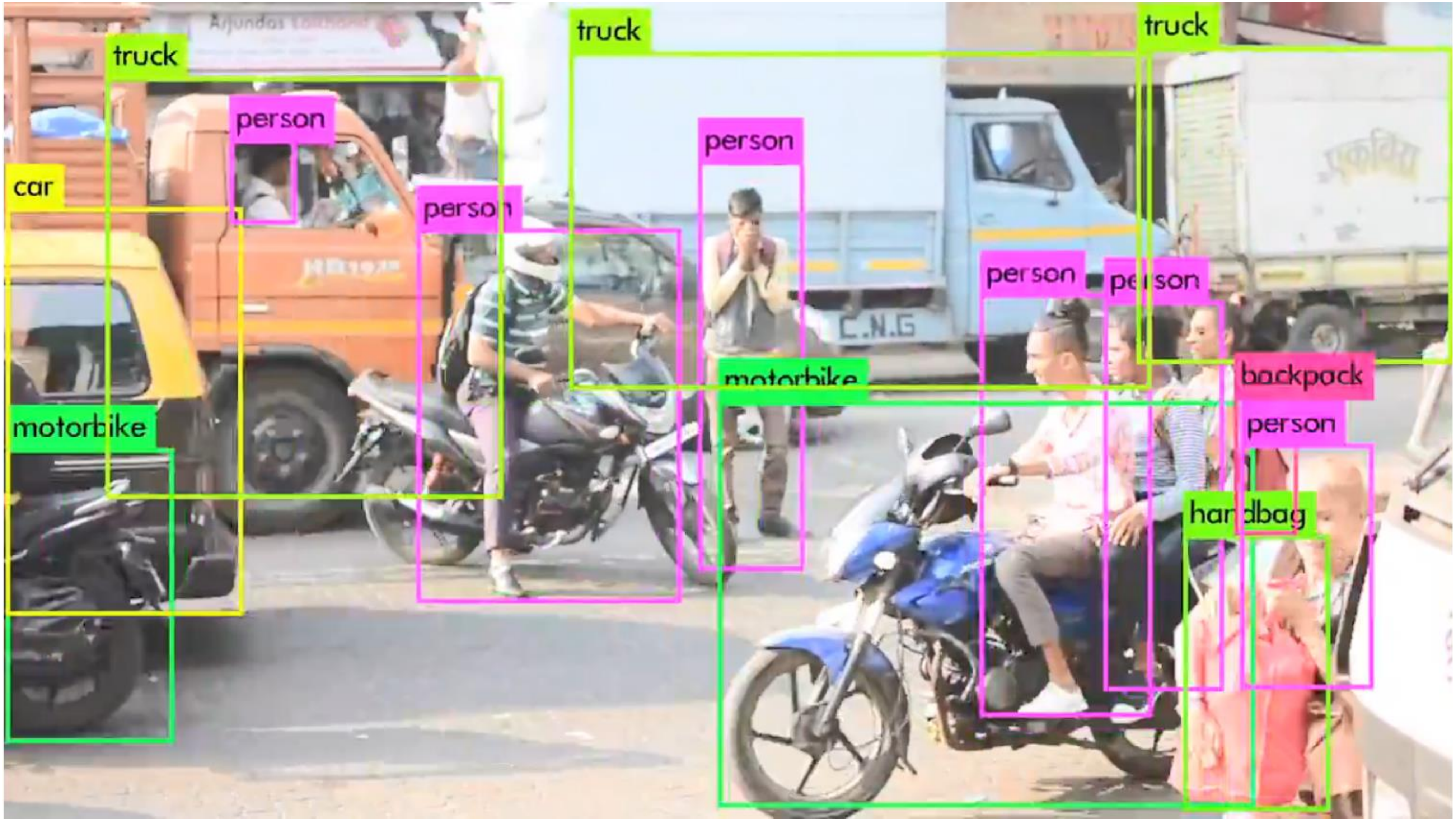
YOLOv3



[Redmond et al., 2018]

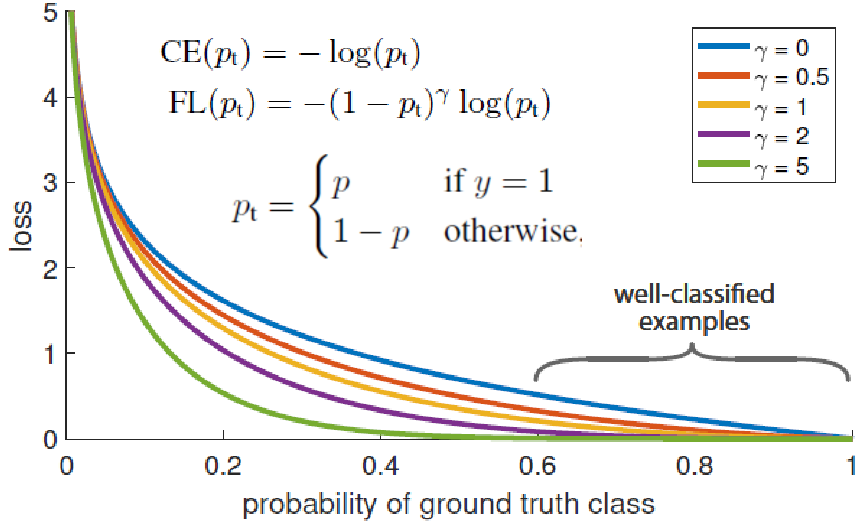
Images from https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b

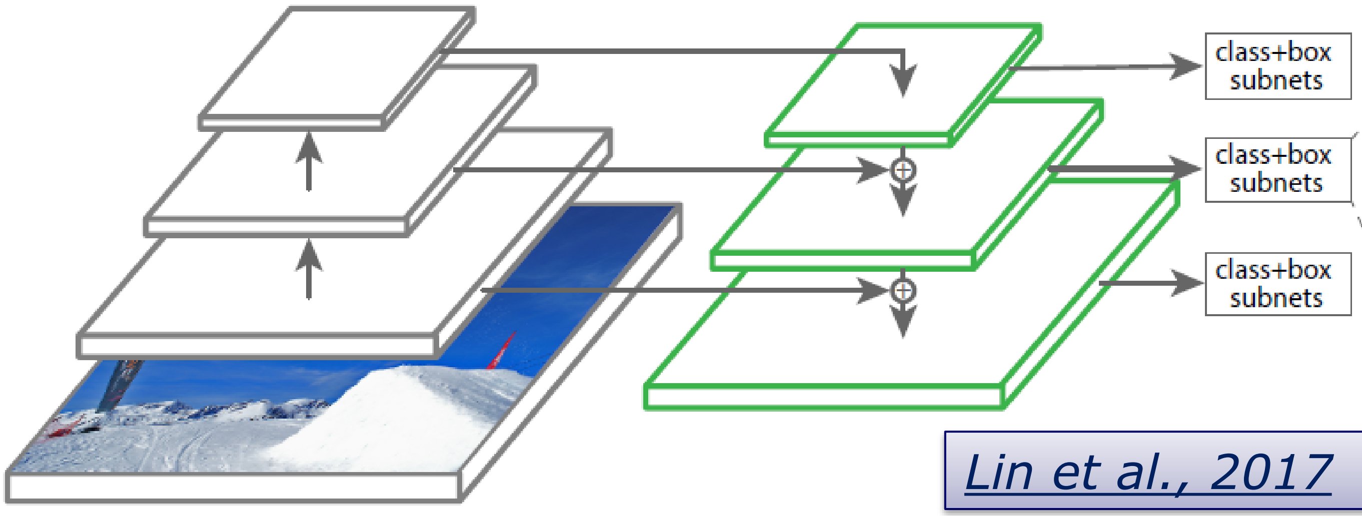
YOLOv3 results

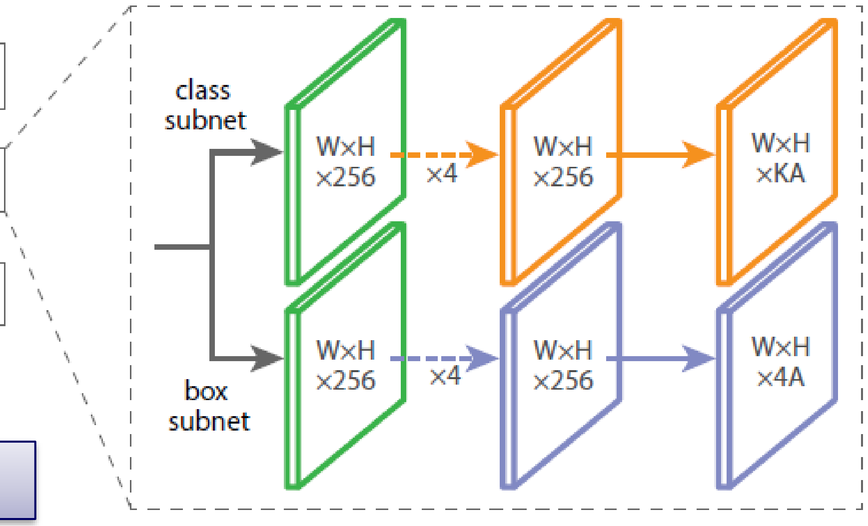


RetinaNet

- Focal Loss for Dense Object Detection
  - Weight loss to deal with class imbalance
  - Dynamically-scaled cross-entropy loss
- RetinaNet – single-stage unified network
  - Backbone: ResNet+FPN
  - Translation invariant anchor boxes (A=9)
  - Classification subnet: small FCN
  - Box regression subnet: class-agnostic rel. offset



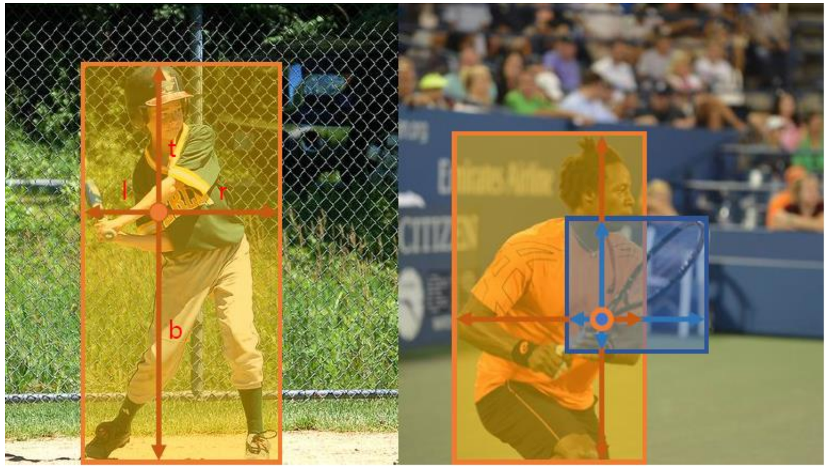


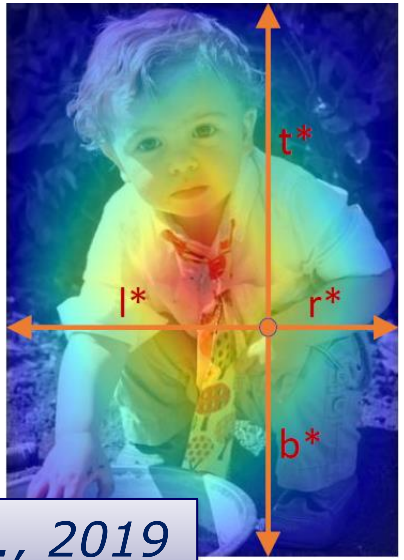


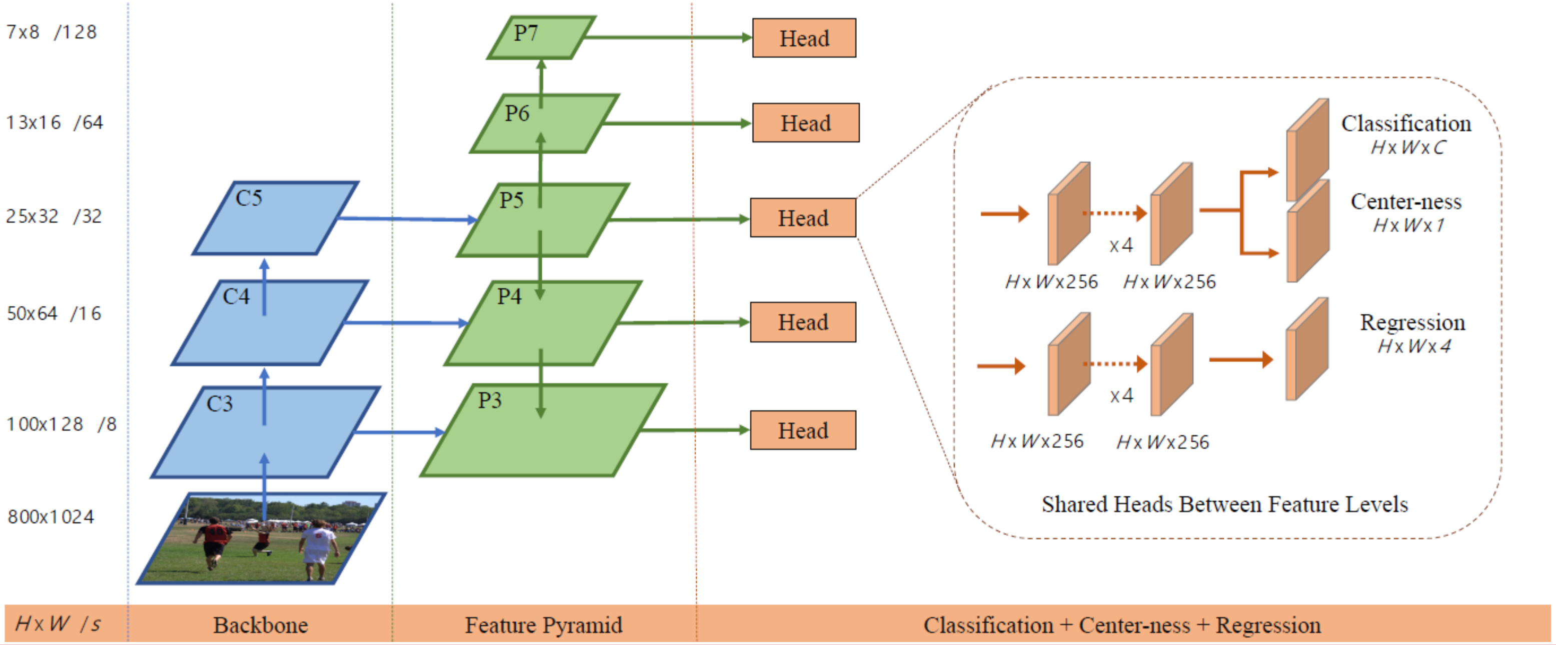
Lin et al., 2017

FCOS: Fully Convolutional One-Stage Object Detection

- Fully convolutional
  - Approaching segmentation methods
- No proposals, no anchor-boxes
- Regressing distances to bounding box
- Multilevel prediction with FPN
- Center-ness to down-weight distant pixels
- Non-maximal suppression

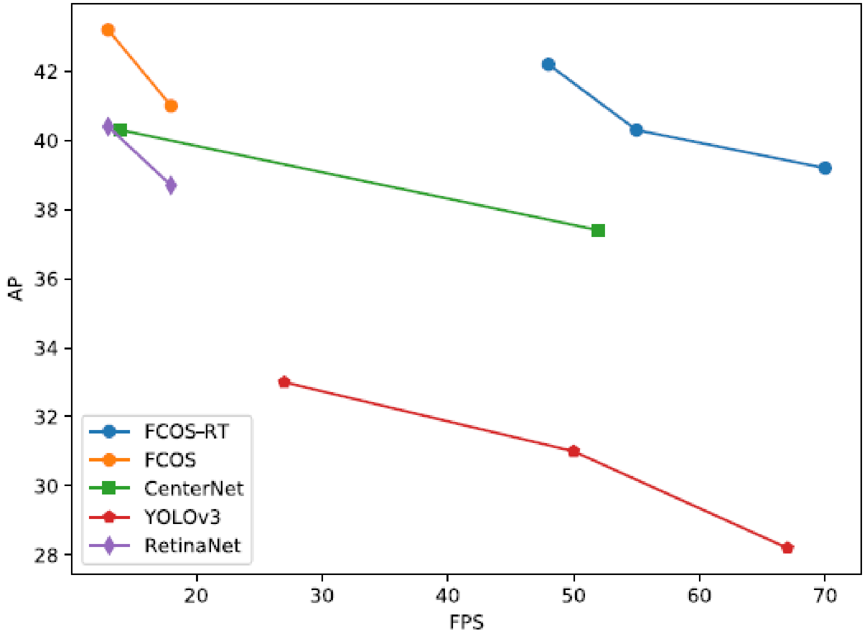






Tian et al., 2019

FCOS results

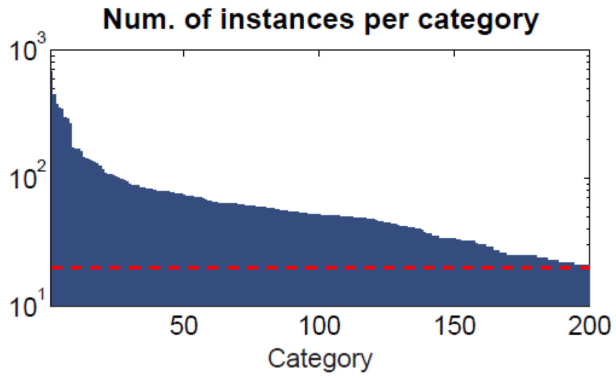


Tian et al., 2019

Tian et al., 2020

Detection of traffic signs

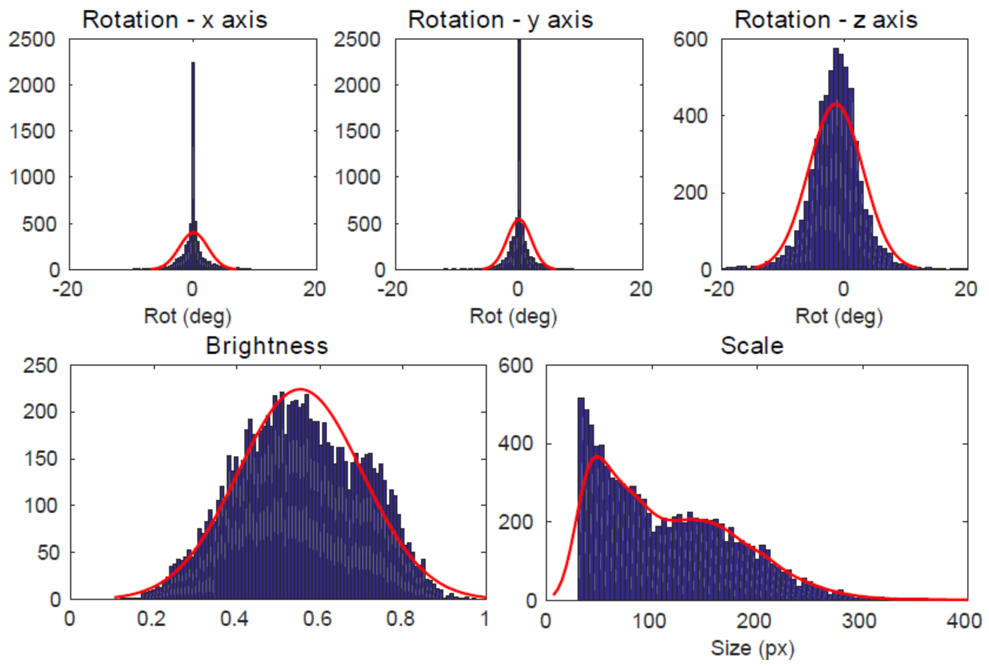
- DFG database
- 200 categories
- 6.957 images
- 13.239 signs

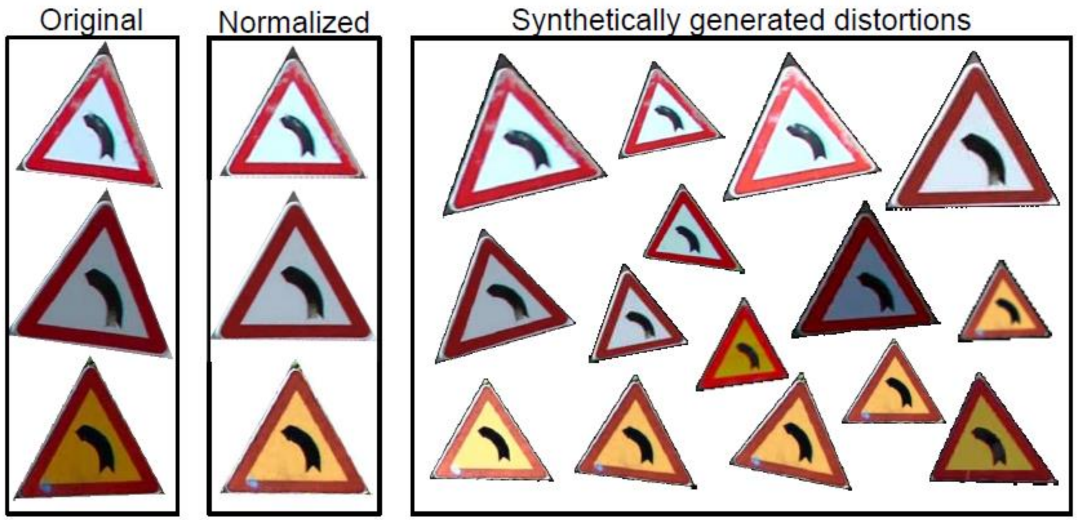


Tabernik & Skočaj, 2020

Detection of traffic signs

- Data augmentation

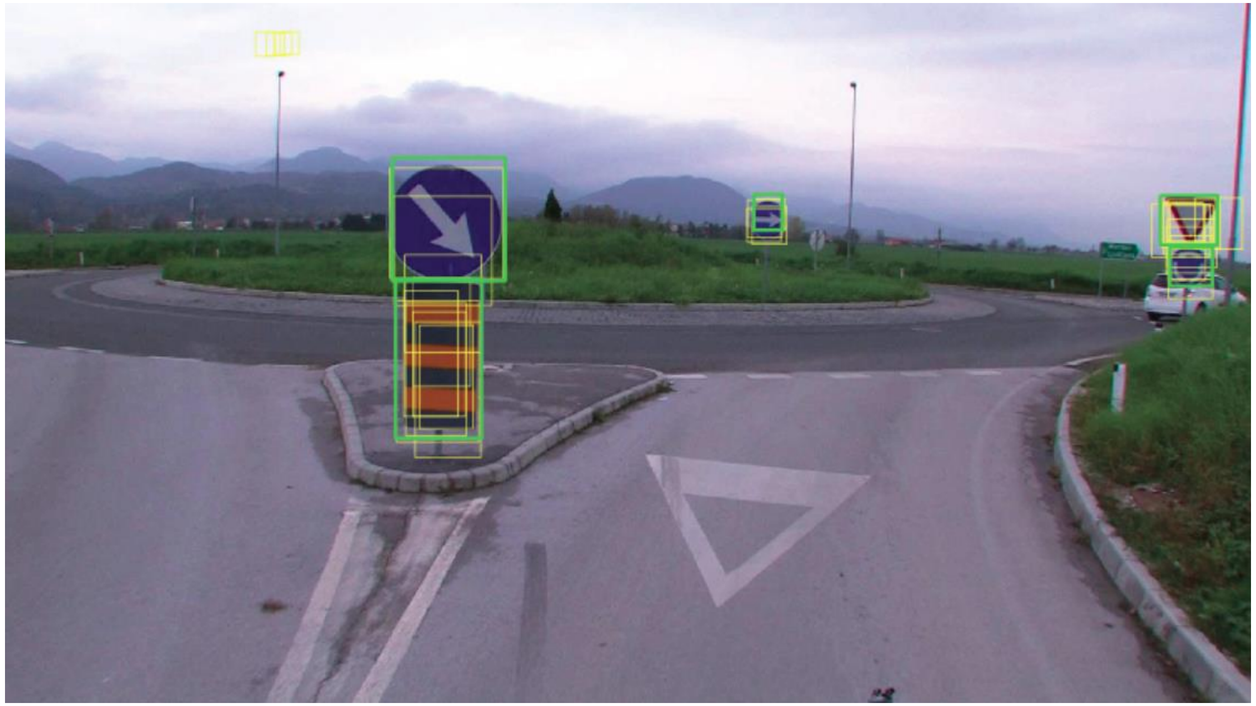


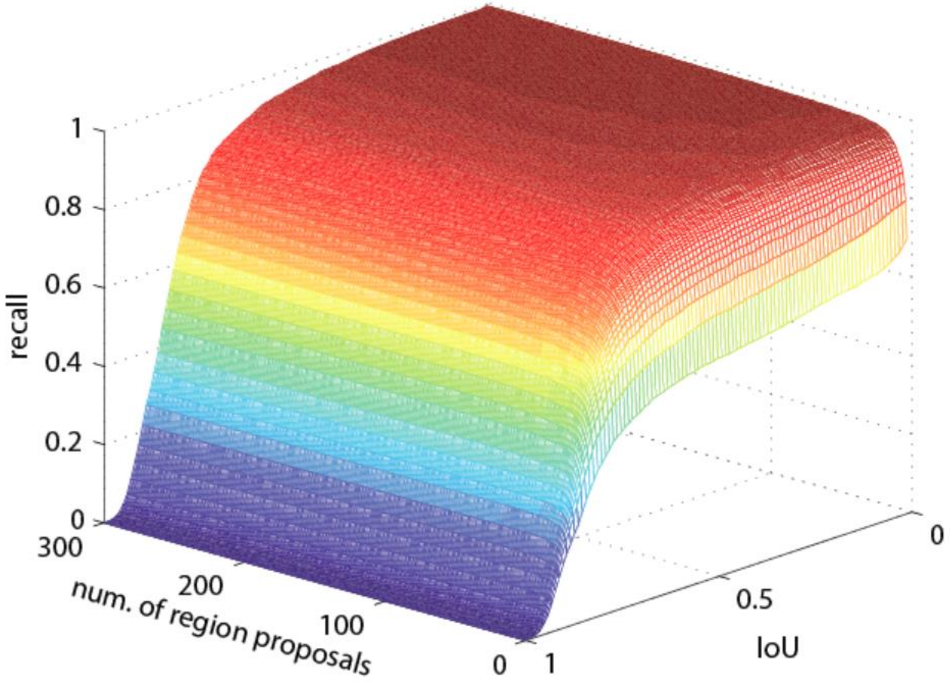


- Mask R-CNN +
  - Online hard-example mining
  - Distribution of selected training samples
  - Sample weighting
  - Adjusting region pass-through during detection

Detection of region proposals

- Top proposals are very good

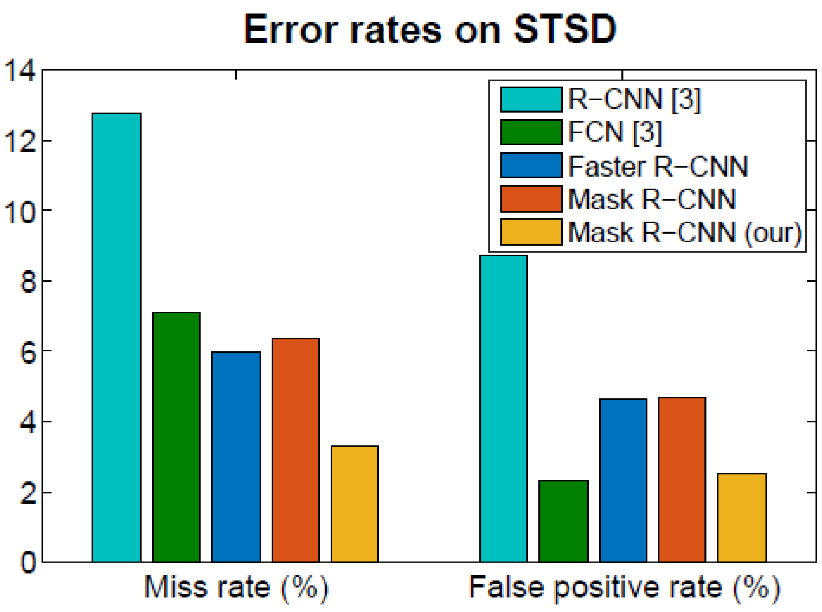




Experimental results

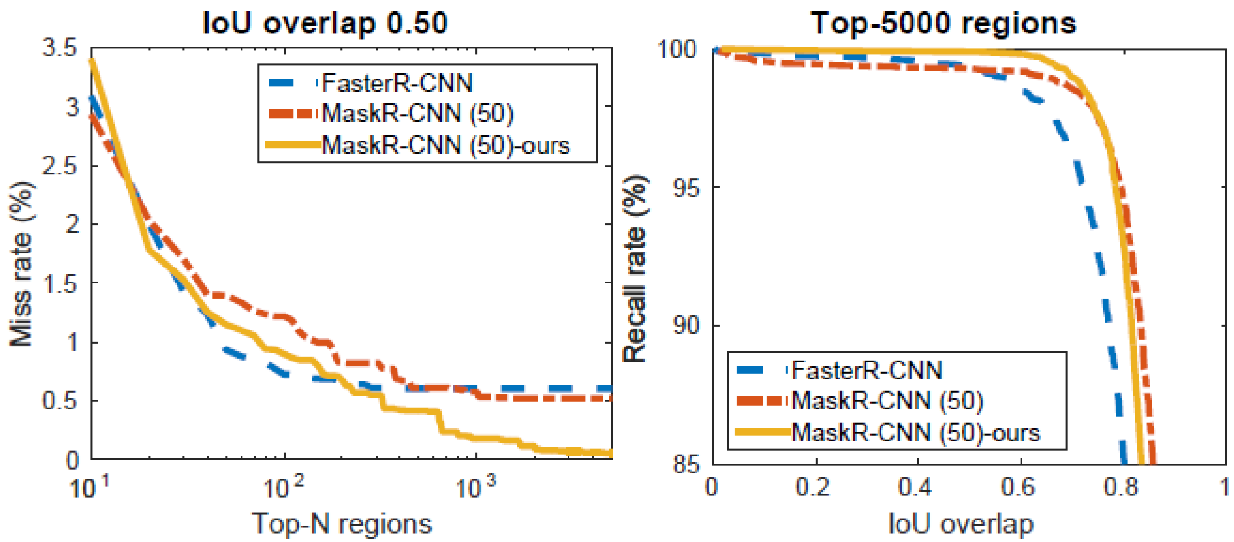
- Swedish traffic sign dataset

Average	R-CNN [6]	FCN [6]	Faster R-CNN	Mask R-CNN (ResNet-50)	
				No adapt.	Adapt. (ours)
Precision	91.2	**97.7**	95.4	95.3	97.5
Recall	87.2	92.9	94.0	93.6	**96.7**
F-measure	88.8	95.0	94.6	93.8	**97.0**
mAP^{50}	/	/	94.3	94.9	**95.2**

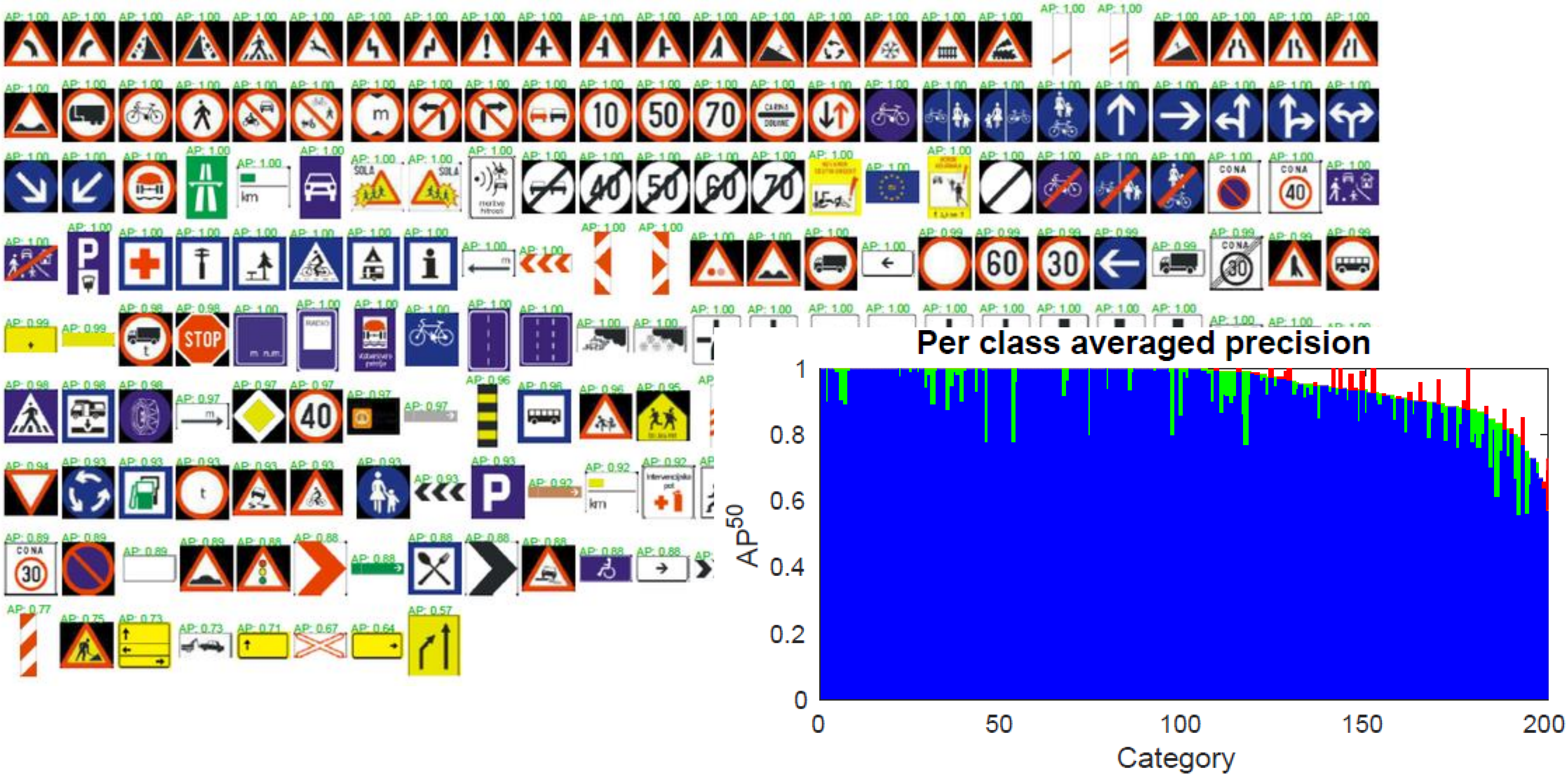


DFG traffic sign dataset

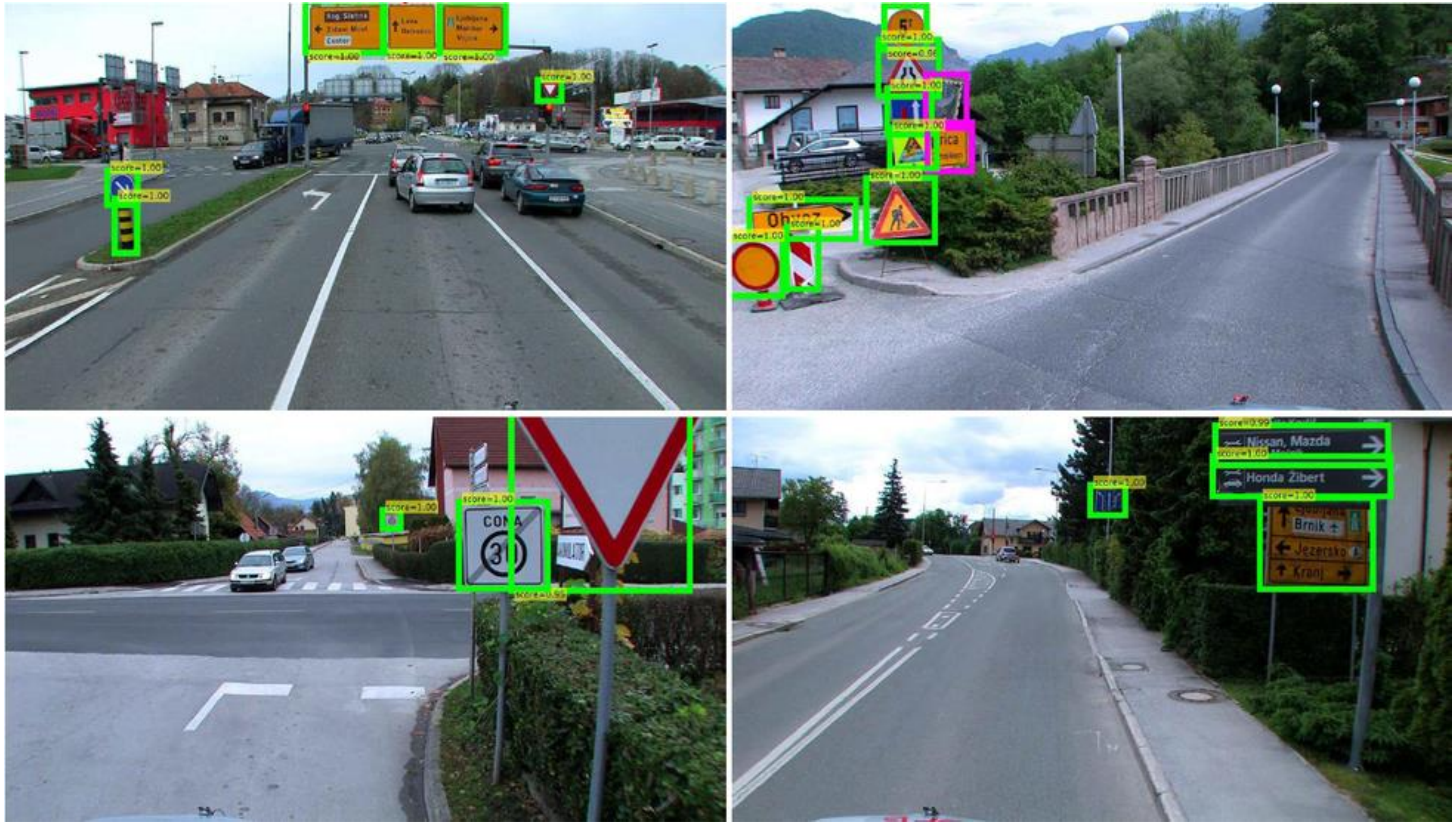
	Faster R-CNN	Mask R-CNN (ResNet-50)		
		No adapt.	With adapt.	With adapt. and data augment.
mAP^{50}	92.4	93.0	95.2	**95.5**
$mAP^{50:95}$	80.4	82.3	82.0	**84.4**
Max recall	93.8	94.6	**96.5**	**96.5**



Experimental results



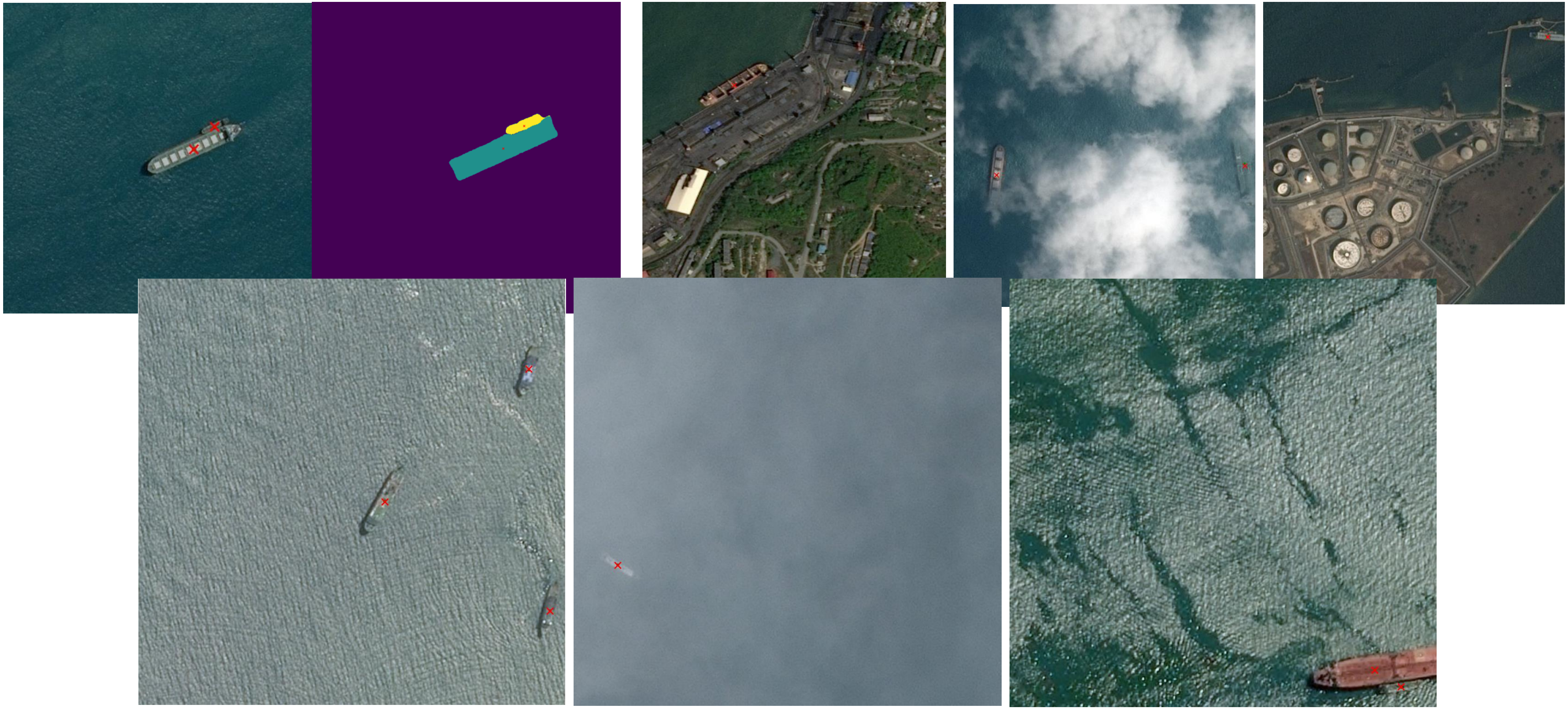
Experimental results



Experimental results

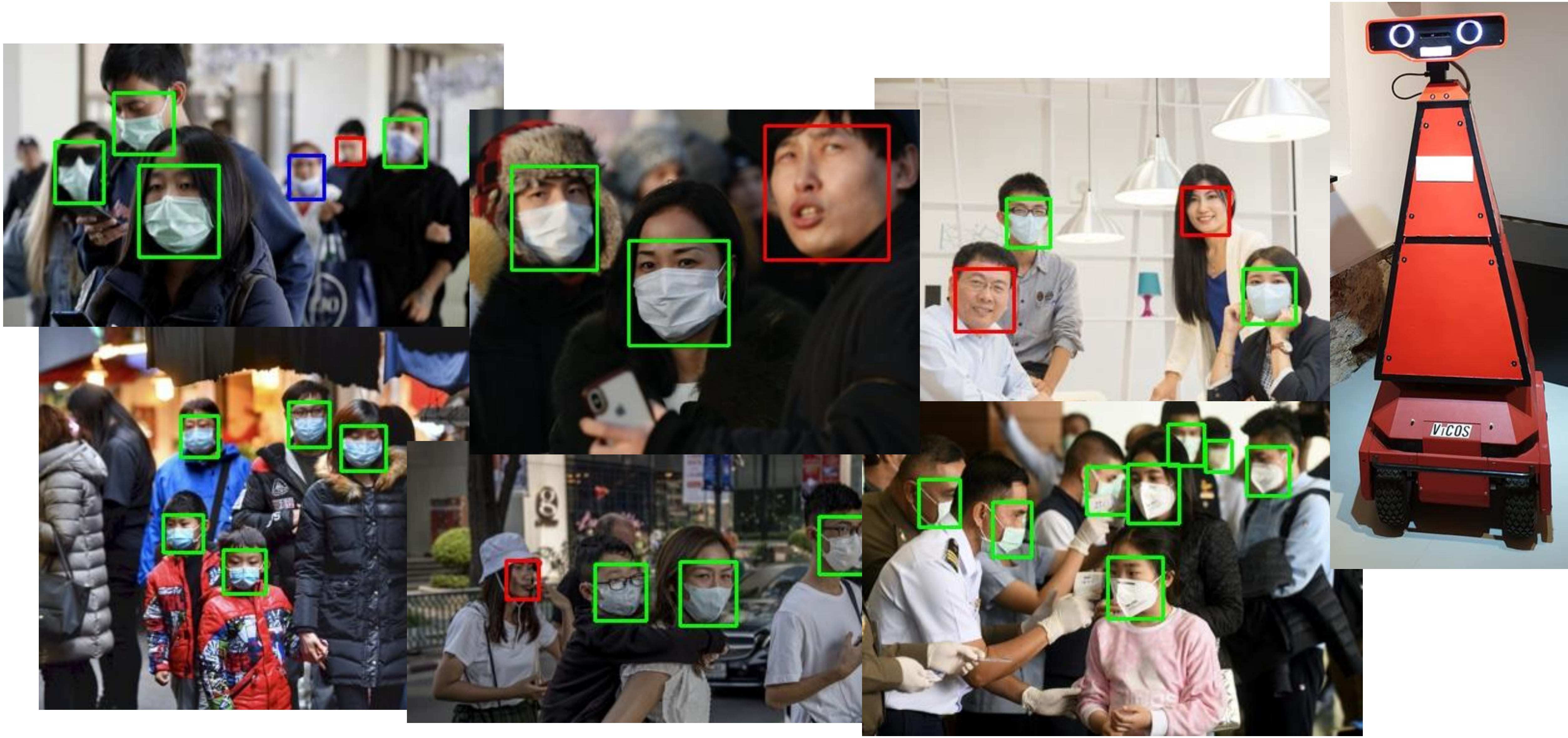
Traffic sign detection

Ship detection

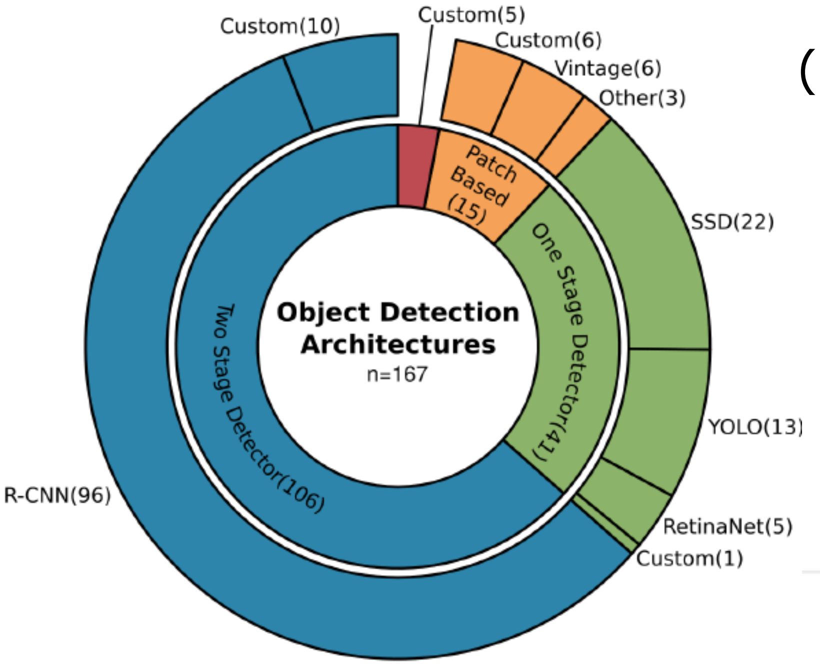


Face detection

Mask-wearing detection



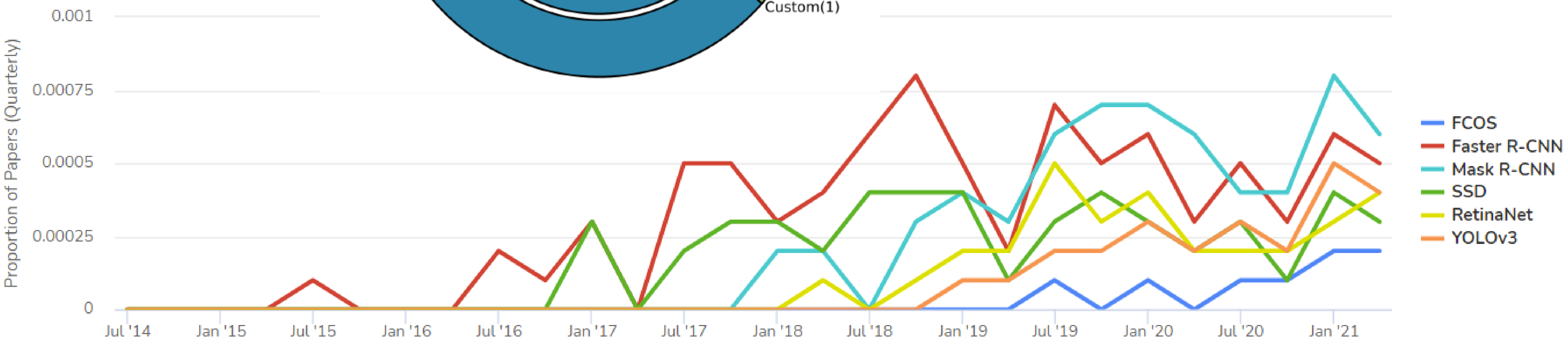
Object detection arhitectures overview



(remote sensing domain)

[Hoeser et. al 2020]

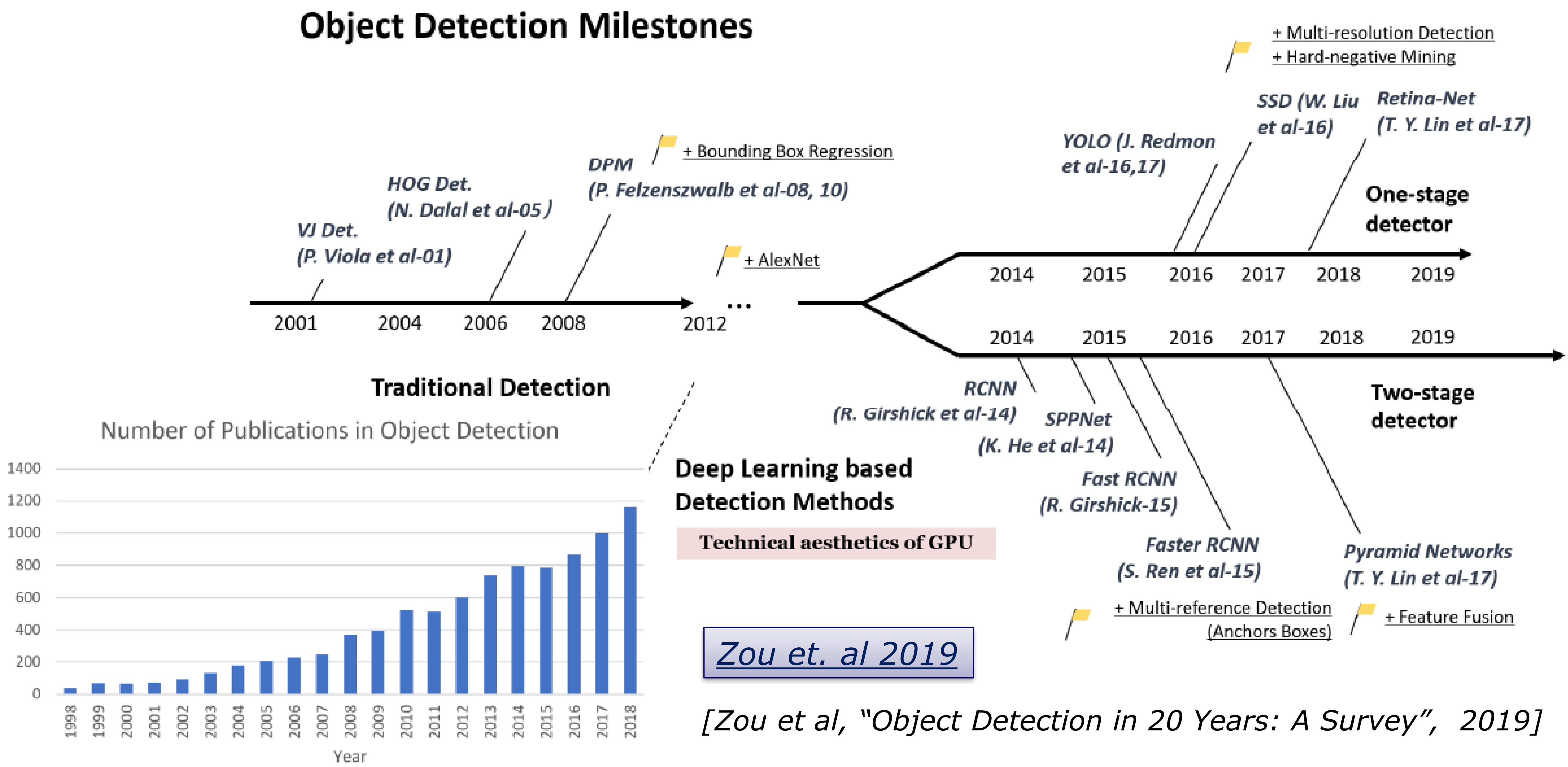
Usage Over Time



This feature is experimental; we are continuously improving our matching algorithm.

[paperswithcode.com, 2021]

Object detection overview



[Zou et. al 2019](#)

[Zou et al, "Object Detection in 20 Years: A Survey", 2019]

Performance of object detectors

- Benchmark datasets
  - Pascal Visual Object Classes (20 classes)
  - ImageNet Large Scale Visual Recognition Challenge (200 classes)
  - MS-COCO (80 classes)
- Metrics
  - Average precision
    - At IoU 0.5
    - Averaged over AP at 0.5:.5:.95
  - mAP: Mean average precision

Zou et. al 2019

Dataset	train		validation		trainv	
	images	objects	images	objects	images	
VOC-2007	2,501	6,301	2,510	6,307	5,011	
VOC-2012	5,717	13,609	5,823	13,841	11,540	
ILSVRC-2014	456,567	478,807	20,121	55,502	476,688	
ILSVRC-2017	456,567	478,807	20,121	55,502	476,688	
MS-COCO-2015	82,783	604,907	40,504	291,875	123,287	
MS-COCO-2018	118,287	860,001	5,000	36,781	123,287	
OID-2018	1,743,042	14,610,229	41,620	204,621	1,784,662	1