Master in
Computer Vision
Barcelona

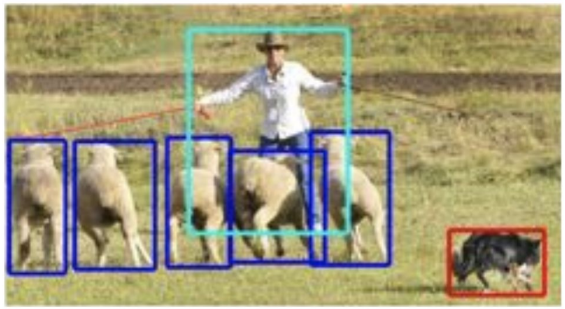
Week 2:
Semantic and Instance Segmentation 2
Visual Recognition

Issam Laradji
issam.laradji@servicenow.com

March 15, 2023

Previously in Visual Recognition

(a) image classification



(b) object detection

(c) semantic segmentation

(d) instance segmentation

Popular Conferences

CVPR: IEEE Conference on Computer Vision and Pattern Recognition

- Premier conference in computer vision
- Held annually

ECCV: European Conference on Computer Vision

- Biennial conference held in Europe

ICCV: International Conference on Computer Vision

- Biennial conference

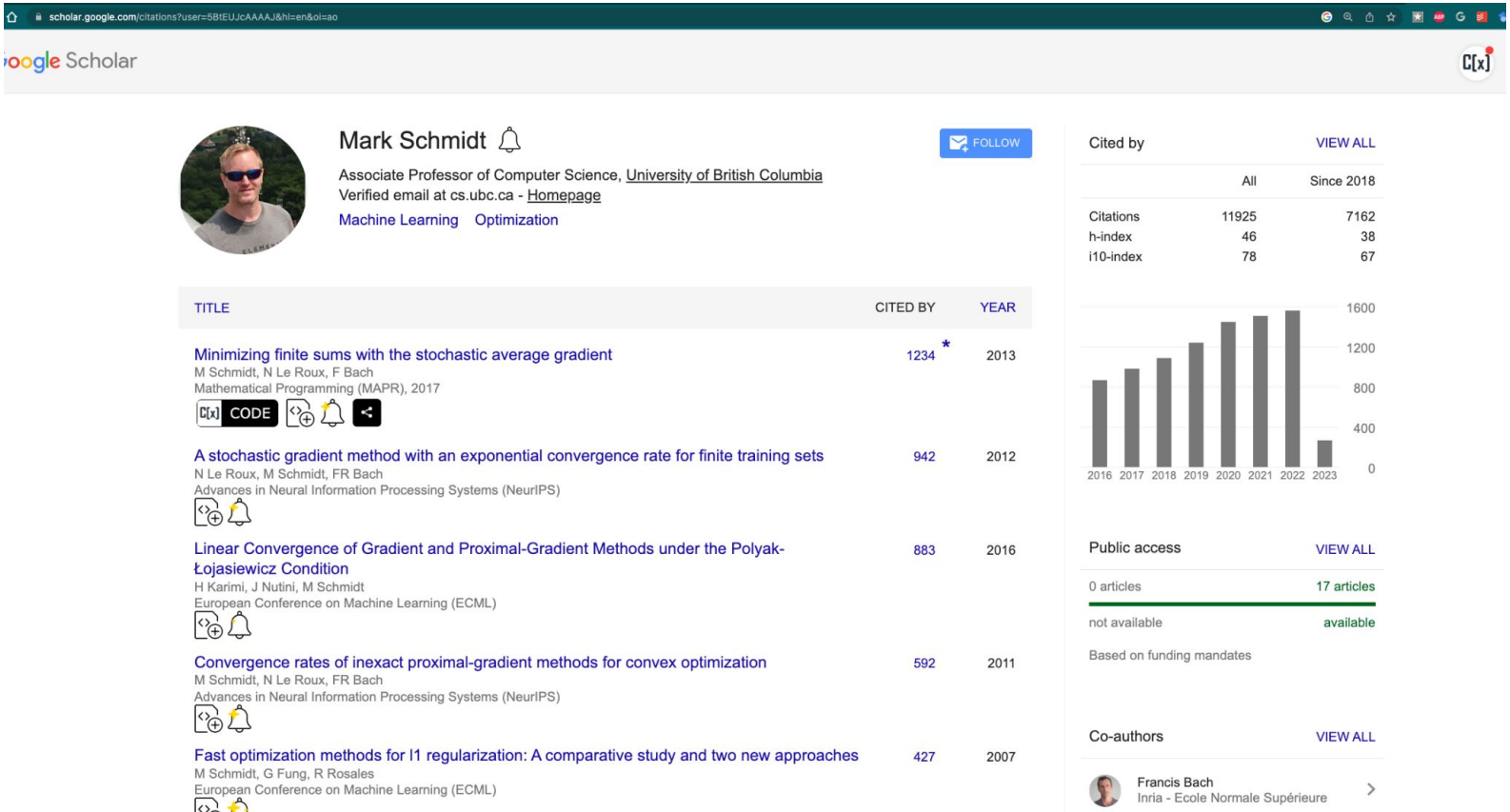
ACCV: Asian Conference on Computer Vision

- Biennial conference focused on computer vision research in the Asia-Pacific region

BMVC: British Machine Vision Conference

- Annual conference held in the UK

Google Scholar



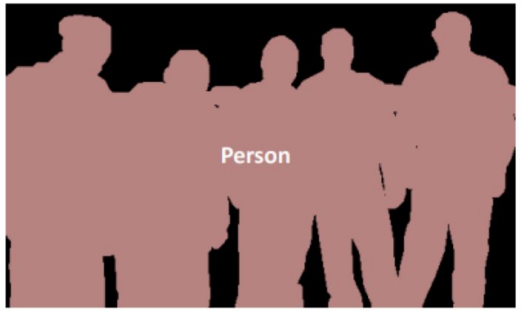
Ask ChatGPT?

Outline

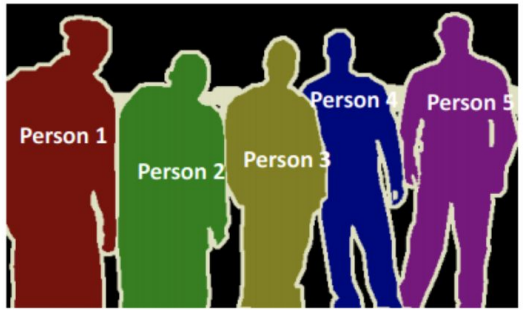
- **Instance segmentation**
- Panoptic segmentation
- Amodal segmentation
- Referring image segmentation
- Current trends and future research

Instance segmentation: Problem statement

- Label regions that fully cover object instances
- Extends object detection with pixel-wise annotations
- Harder than semantic segmentation
- Current state-of-the-art:
  - object proposal + classification



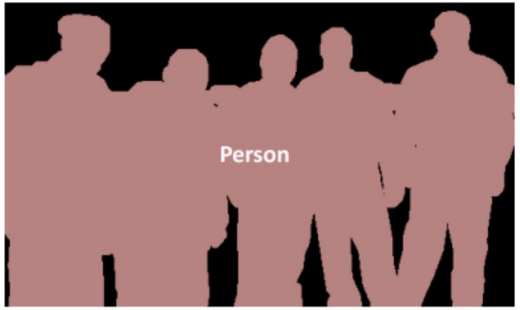
Semantic Segmentation



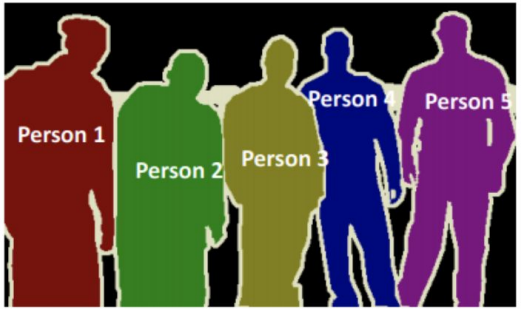
Instance Segmentation

Instance segmentation: Problem statement

- Harder than semantic segmentation:
  - The representation of the problem is harder
  - It falls into the category of structured prediction
  - Naive representation does not work:
    - do not train on instance IDs
  - There is no clear winner yet
  - Several ways to deal with the problem:
    - Candidate proposal, bounding box detection and pixel-mask refinement
    - Attention models and RNNs
    - Partition space and metric learning
    - Transformers



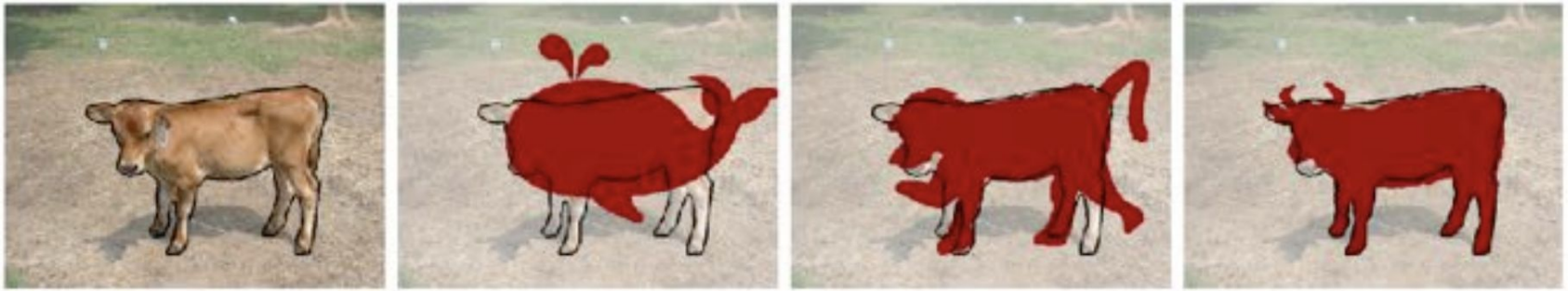
Semantic Segmentation



Instance Segmentation

Instance segmentation: Metrics

- Same as in Object Detection...
  - IoU between ground truth mask and predicted mask
  - mean average precision (AP)
  - mean average recall (AR)
- … but instead of bounding boxes, using pixel-wise masks

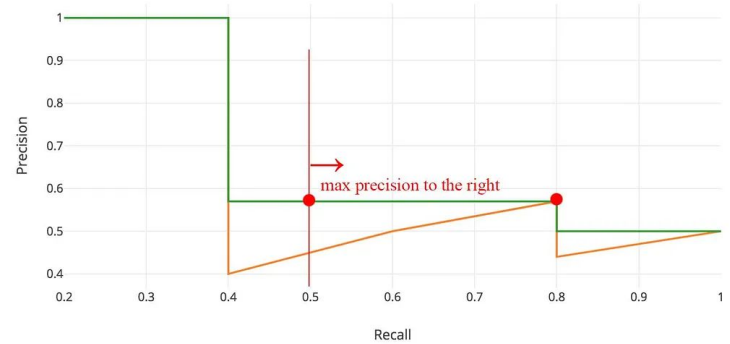


(a) ground truth (b) IoU = 0.554 (c) IoU = 0.703 (d) IoU = 0.910

Figure from Krahenbuhl and Koltun. Geodesic Object Proposals. ECCV14

Instance segmentation: Metrics

Rank	Correct?	Precision	Recall
1	True	1.0	0.2
2	True	1.0	0.4
3	False	0.67	0.4
4	False	0.5	0.4
5	False	0.4	0.4
6	True	0.5	0.6
7	True	0.57	0.8
8	False	0.5	0.8
9	False	0.44	0.8
10	True	0.5	1.0



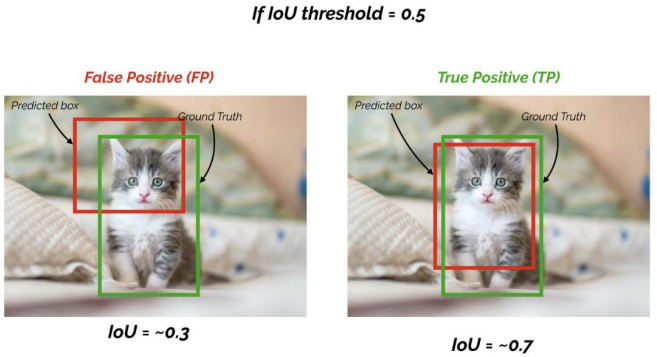
Precision is the proportion of TP = 2/3 = 0.67.

Recall is the proportion of TP out of the possible positives = 2/5 = 0.4.

Average Precision (AP) is finding the area under the precision-recall curve above.

Source: https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173

Instance segmentation: Metrics

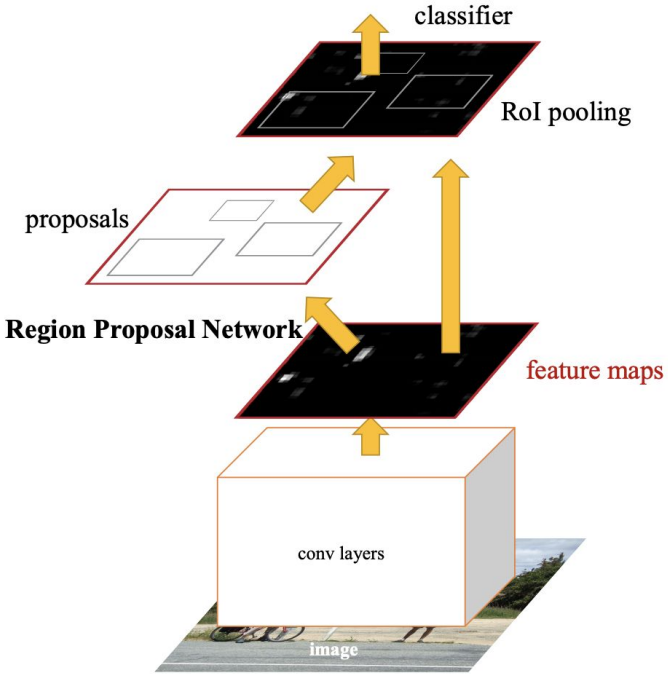


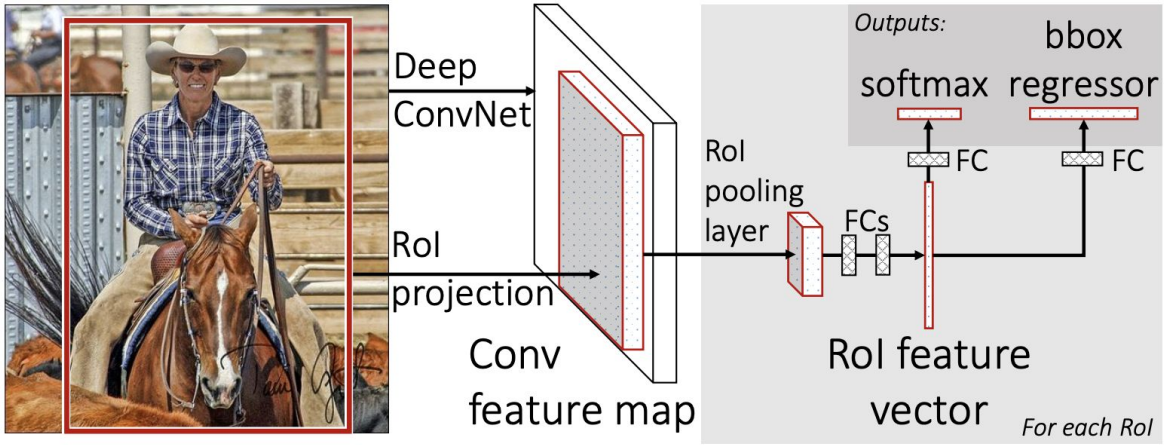
- Do not implement your own!
  - Use **pycocotools**
- For the COCO competition,
  - AP is the average over 10 IoU levels on 80 categories
    - AP@[.50:.05:.95]: start from 0.5 to 0.95 with a step size of 0.05

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods							
Faster R-CNN+++ [5]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [8]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [6]	Inception-ResNet-v2 [21]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [20]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
One-stage methods							
YOLOv2 [15]	DarkNet-19 [15]	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [11, 3]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [3]	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet [9]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet [9]	ResNeXt-101-FPN	**40.8**	**61.1**	**44.1**	**24.1**	**44.2**	51.2
YOLOv3 608 × 608	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9

Instance segmentation: Region-based techniques

- Review **SOTA** Object Detection: Fast R-CNN & Faster R-CNN



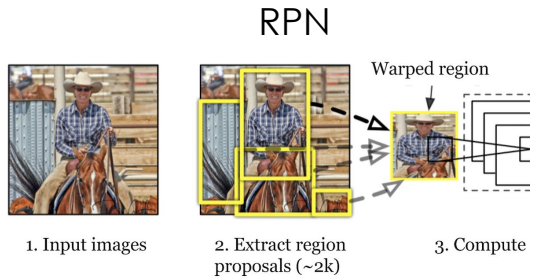


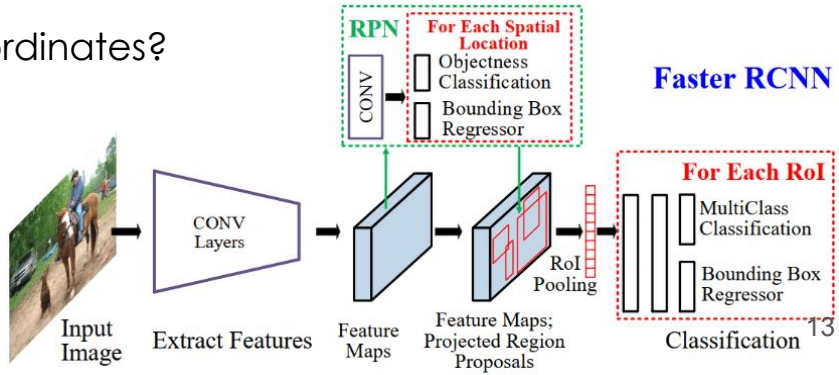
Instance segmentation: Region-based techniques

- Review **SOTA** Object Detection: Fast R-CNN & Faster R-CNN

One network, four losses (**TPAMI** version):

- **RPN** classification loss
  - is the proposal an object or not?
- **RPN** regression loss
  - how much can we change the (x,y) coordinates?
- **Fast(er) R-CNN** classification loss
  - which class is the proposal?
- **Fast(er) R-CNN** regression
  - how much can we change the (x,y) coordinates?





Instance segmentation: Region-based techniques

- Mask R-CNN [ICCV2017] (>5k citations)

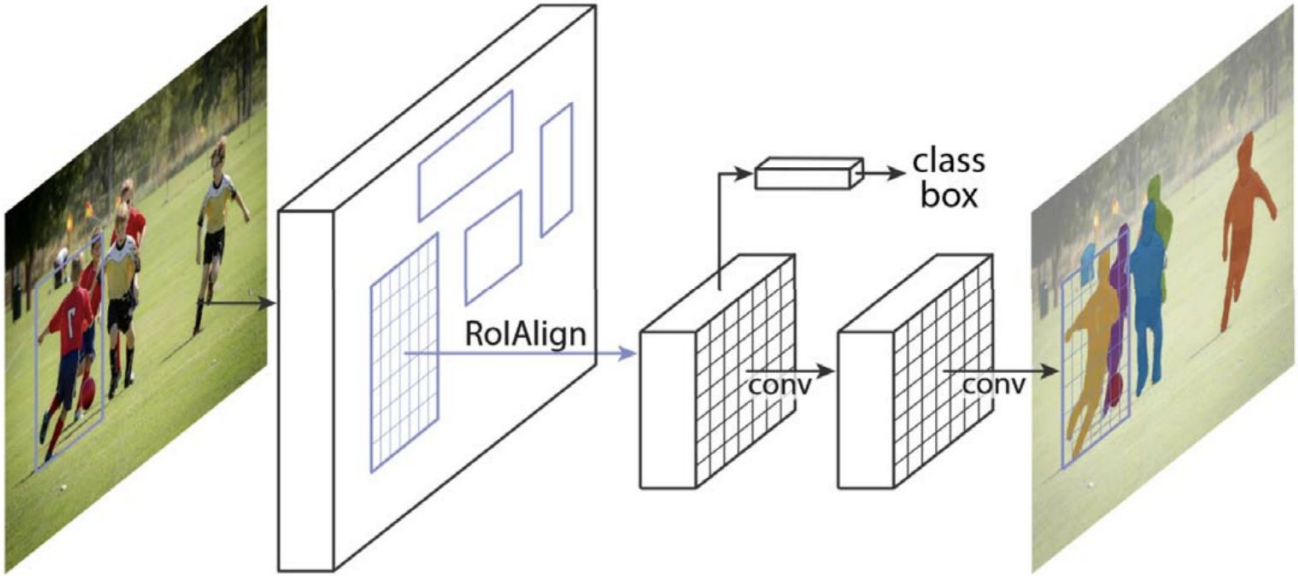
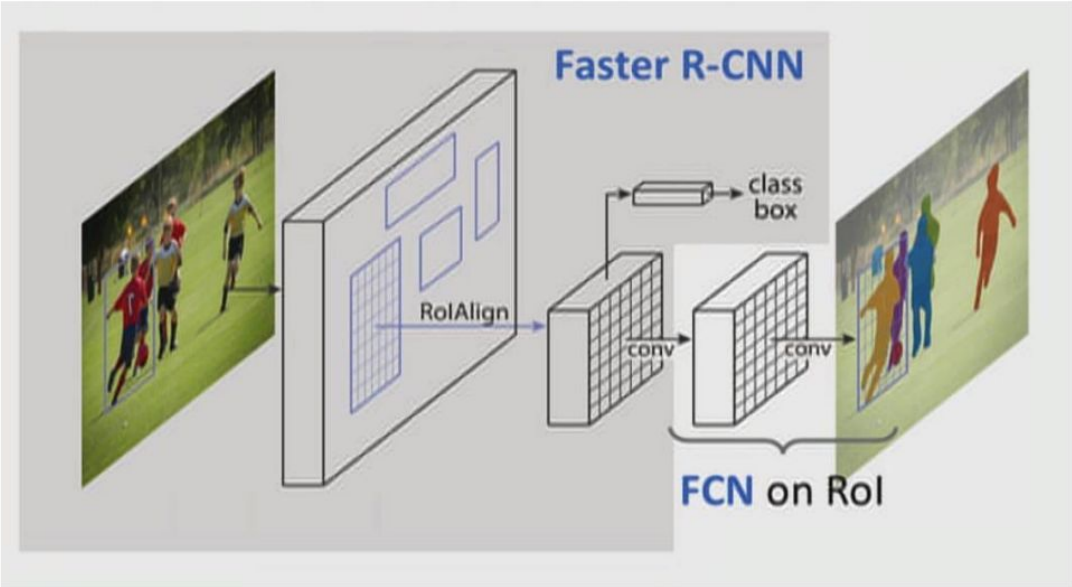
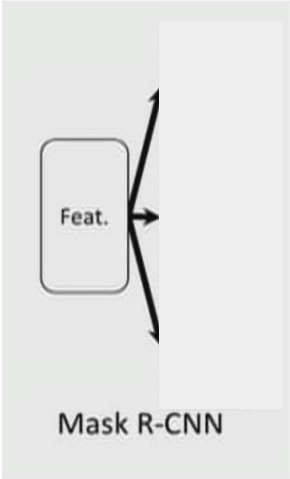


Image sourced from He, Kaiming, et al. "Mask R-CNN."

Instance segmentation: Region-based techniques

- Mask R-CNN [ICCV2017] (>5k citations)

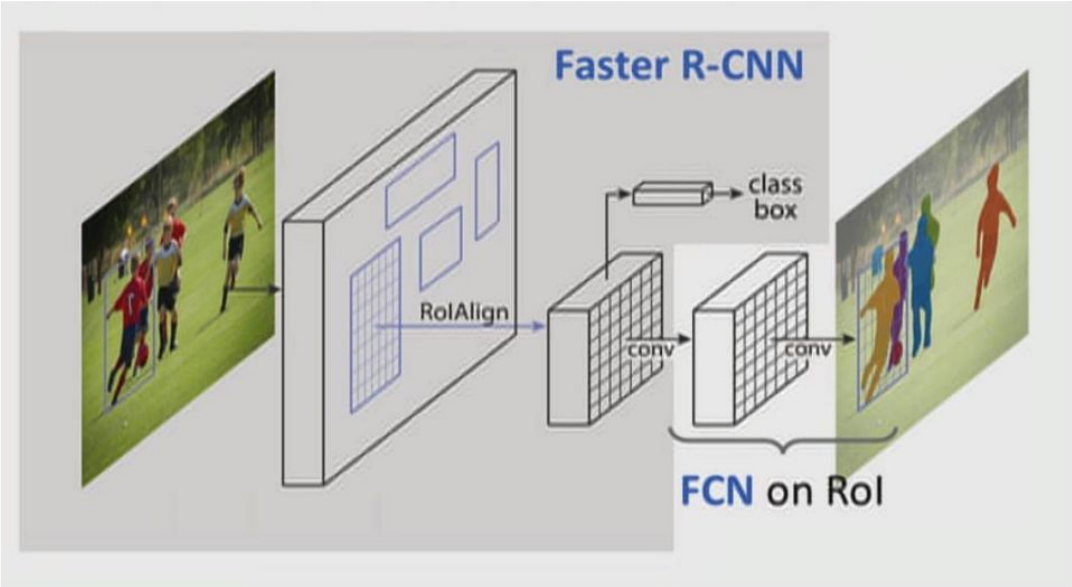


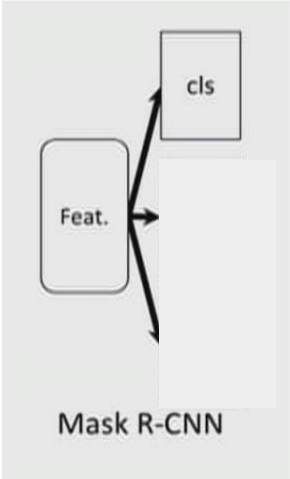


Loss Functions

Instance segmentation: Region-based techniques

- Mask R-CNN [ICCV2017] (>5k citations)

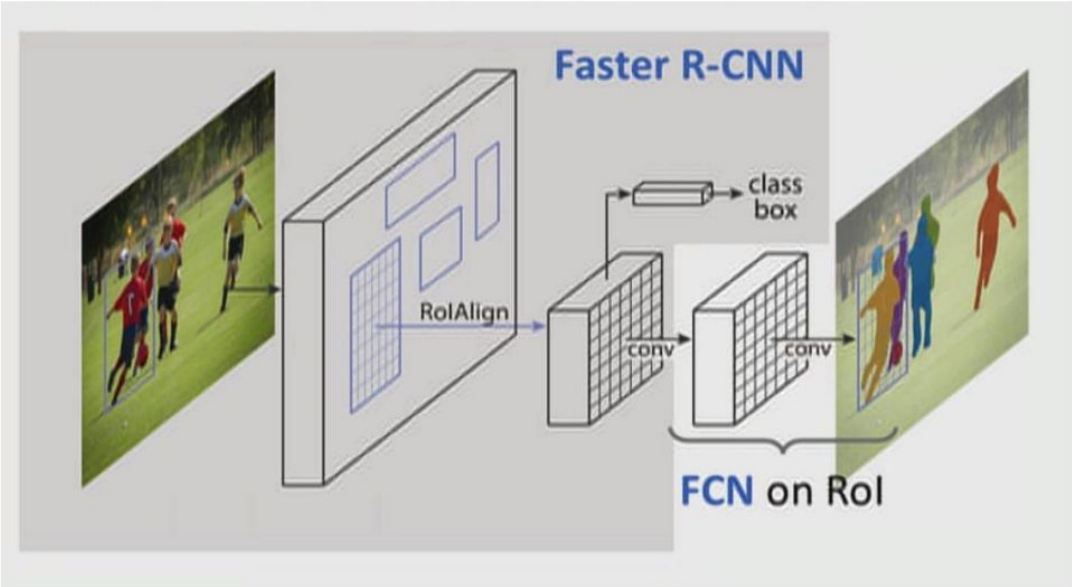


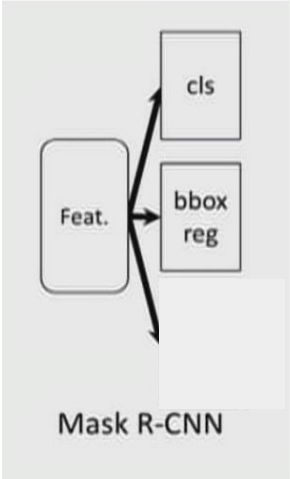


Loss Functions

Instance segmentation: Region-based techniques

- Mask R-CNN [ICCV2017] (>5k citations)

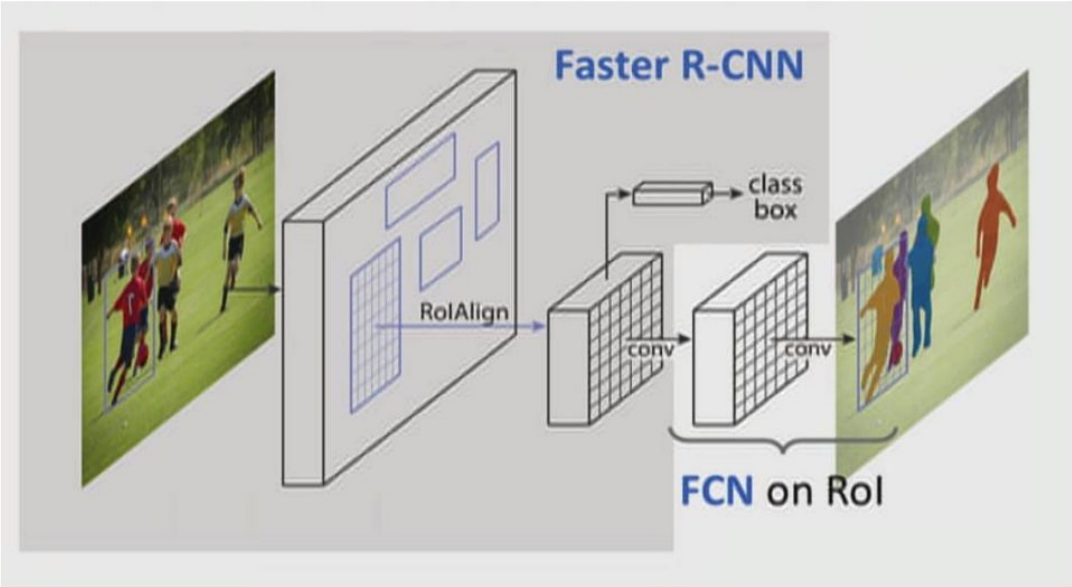


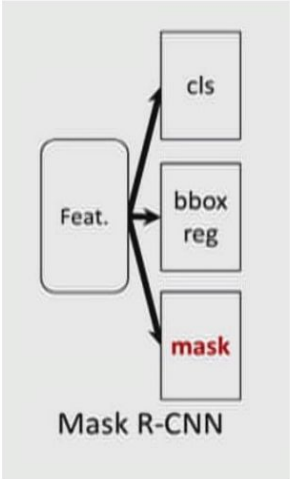


Loss Functions

Instance segmentation: Region-based techniques

- Mask R-CNN [ICCV2017] (>5k citations)

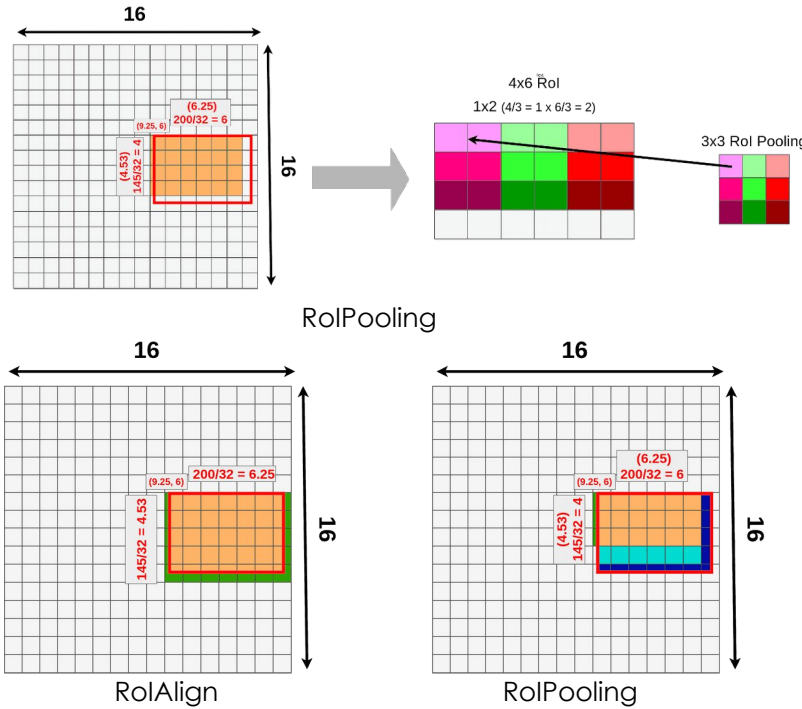




Loss Functions

Instance segmentation: Region-based techniques

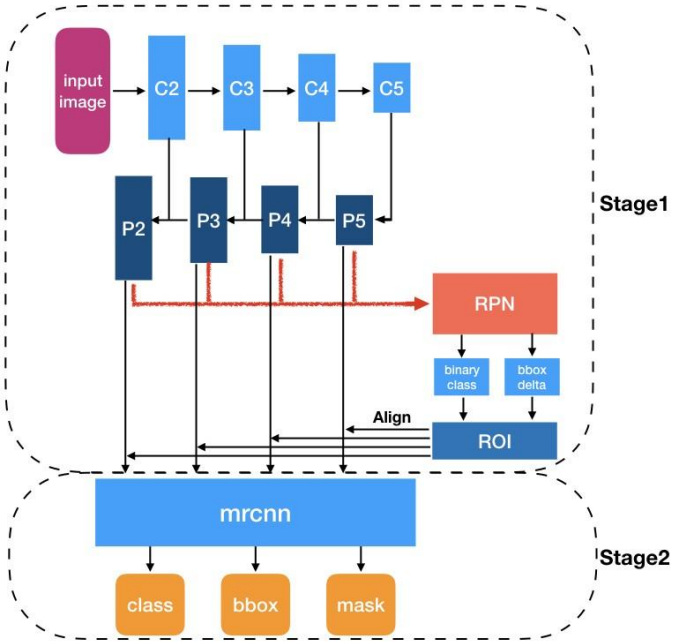
- Mask R-CNN [ICCV2017] (>5k citations)
- Extends Faster RCNN in 2 main ways
  - **Mask branch:** generates pixel-level instance segmentation masks for each detected object (Uses Pixel-level Cross Entropy loss)
  - **RoI Align:**
    - no quantization compared to ROIPooling, uses bilinear interpolation
    - more accurate feature alignment between the feature map and RoIs.
    - blue (lost info.)
    - green (gained info.)



Source: https://erdem.pl/2020/02/understanding-region-of-interest-part-2-ro-i-align

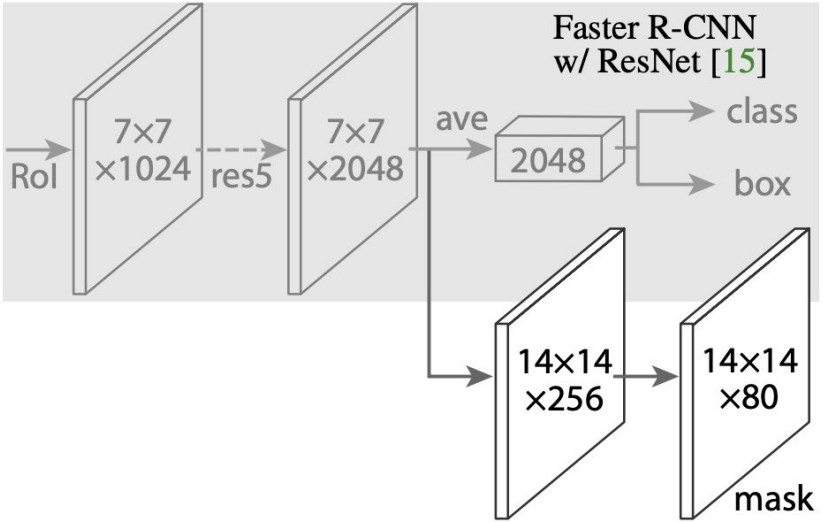
Instance segmentation: Region-based techniques

- Mask R-CNN [ICCV2017] (>5k citations)



Instance segmentation: Region-based techniques

- Mask R-CNN [ICCV2017] (>5k citations)
  - Mask branch: FCN that outputs a binary mask for each ROI
    - Mask represented as C x m x m (C classes, m x m ROI) -> one mask per class
    - Binary loss on the mask: predict 1 where pixel if belongs to object and 0 otherwise



	AP	AP_{50}	AP_{75}
softmax	24.8	44.1	25.1
sigmoid	**30.3**	**51.2**	**31.5**
	+5.5	*+7.1*	*+6.4*

(b) **Multinomial *vs*. Independent Masks** (ResNet-50-C4): *Decoupling* via per-class binary masks (sigmoid) gives large gains over multinomial masks (softmax).

Instance segmentation: Region-based techniques

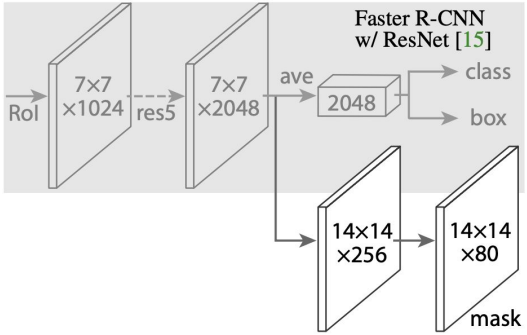
- Mask R-CNN [ICCV2017] (>5k citations)
  - Mask prediction requires estimation of pixel mask:
    - ROIPool (used in Fast(er) R-CNN): it contains 2 quantization steps
      - ok for object detection, but poor for mask segmentation
    - ROIPool -> ROIAlign: it uses bilinear interpolation to avoid quantization

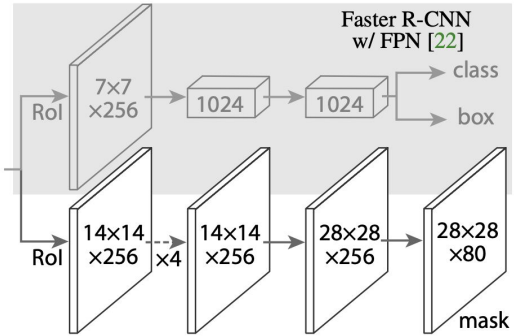
	align?	bilinear?	agg.	AP	AP_{50}	AP_{75}
RoIPool [12]			max	26.9	48.8	26.4
RoIWarp [10]		✓	max	27.2	49.2	27.1
		✓	ave	27.1	48.9	27.1
RoIAlign	✓	✓	max	**30.2**	**51.0**	**31.8**
	✓	✓	ave	**30.3**	**51.2**	**31.5**

(c) **RoIAlign** (ResNet-50-C4): Mask results with various RoI layers. Our RoIAlign layer improves AP by ~3 points and AP_{75} by ~5 points. Using proper alignment is the only factor that contributes to the large gap between RoI layers.

Instance segmentation: Region-based techniques

- Mask R-CNN [ICCV2017] (>5k citations)





net-depth-features	AP	AP_{50}	AP_{75}
ResNet-50-C4	30.3	51.2	31.5
ResNet-101-C4	32.7	54.2	34.3
ResNet-50-FPN	33.6	55.2	35.3
ResNet-101-FPN	35.4	57.3	37.5
ResNeXt-101-FPN	**36.7**	**59.5**	**38.9**

(a) **Backbone Architecture**: Better backbones bring expected gains: deeper networks do better, FPN outperforms C4 features, and ResNeXt improves on ResNet.

Instance segmentation: Region-based techniques

- Mask R-CNN [ICCV2017] (>5k citations)

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [7]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [21] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [21] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**	**16.9**	**39.9**	**53.5**

Table 1. **Instance segmentation** *mask* AP on COCO `test-dev`. MNC [7] and FCIS [21] are the winners of the COCO 2015 and 2016 segmentation challenges, respectively. Without bells and whistles, Mask R-CNN outperforms the more complex FCIS+++, which includes multi-scale train/test, horizontal flip test, and OHEM [30]. All entries are *single-model* results.

Instance segmentation: Region-based techniques

- Mask R-CNN [ICCV2017] (>5k citations)
  - ICCV2017: http://openaccess.thecvf.com/content_ICCV_2017/papers/He_Mask_R-CNN_ICCV_2017_paper.pdf
  - Github repository: https://github.com/facebookresearch/detectron2
  - Two-Stage instance segmentation

Instance segmentation: Region-based techniques

- Path Aggregation Network for Instance Segmentation (PA-Net) [CVPR2018]
  - **Idea**:
    - Features in low levels are helpful for large instance identification
  - **Findings**:
    - Mask R-CNN drawbacks
      - Long path from low-level structure to topmost features
      - Increasing difficulty to access accurate localization information
  - **Goal**: boosting information flow by enhancing the entire feature hierarchy
    - Accurate localization signals in lower layers
    - Bottom-up path augmentation
      - shortens the information path between lower layers and topmost feature

Instance segmentation: Region-based techniques

- Path Aggregation Network for Instance Segmentation (PA-Net) [CVPR2018]
  - **Shortcut (green):** less than 10 layers from low-level structure to topmost features (finer grain)
  - **Longpath (red):** more than 100 layers from low-level structure to topmost features (finer grain)

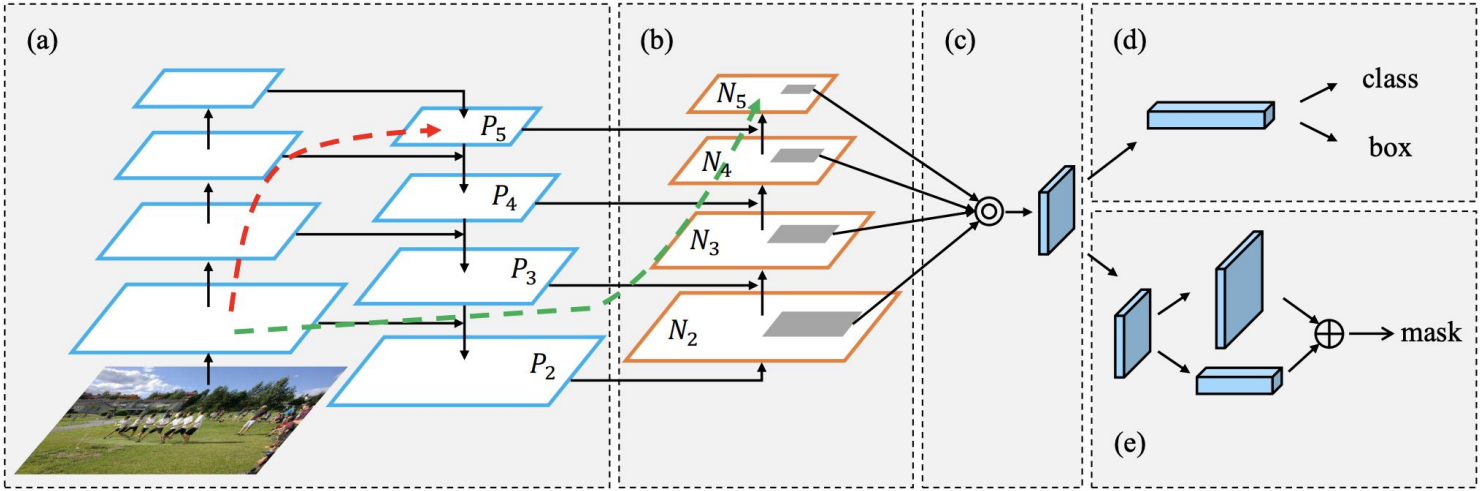


Figure 1. Illustration of our framework. (a) FPN backbone. (b) Bottom-up path augmentation. (c) Adaptive feature pooling. (d) Box branch. (e) Fully-connected fusion. Note that we omit channel dimension of feature maps in (a) and (b) for brevity.

Instance segmentation: Region-based techniques

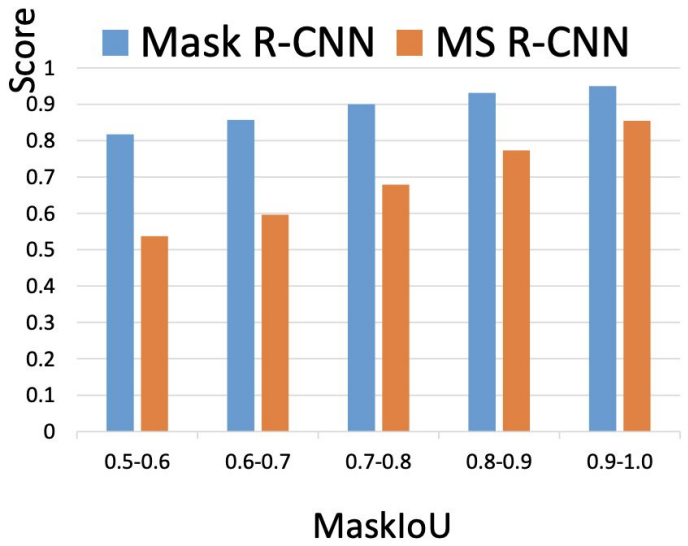
- Path Aggregation Network for Instance Segmentation (PA-Net) (>700 citations) [CVPR2018]
  - http://openaccess.thecvf.com/content_cvpr_2018/papers/Liu_Path_Aggregation_Network_CVPR_2018_paper.pdf
  - Two-Stage instance segmentation

Instance segmentation: Region-based techniques

- Mask Scoring R-CNN (MS-RCNN) [CVPR2019]
  - **Mask R-CNN problem**:
    - Confidence of instance classification is used as mask quality score
    - Mask quality (IoU between predicted mask and GT) is usually not well correlated with classification score
  - **Idea in Mask Scoring R-CNN:**
    - Include a network block to **learn the quality of the predicted instance masks**
    - Mask IoU regression based on:
      - Instance feature
      - Predicted mask
    - Mask scoring strategy calibrates the misalignment between mask quality and mask score

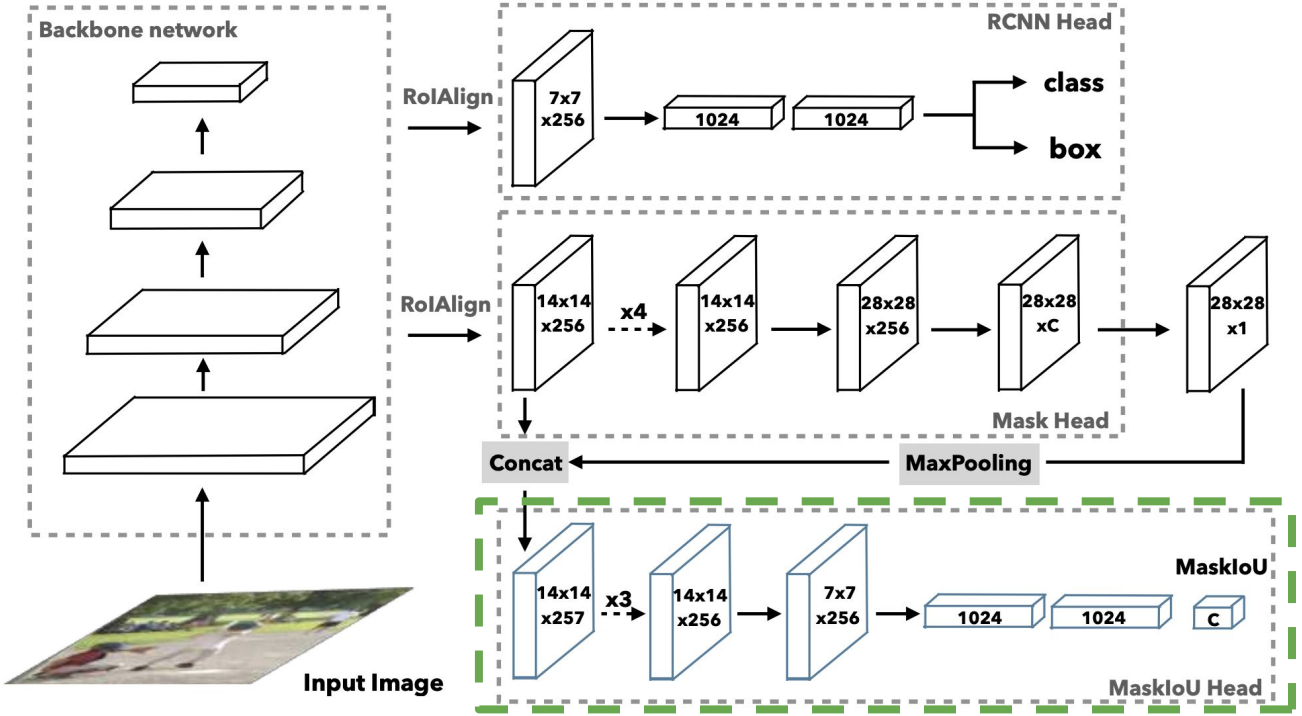
Instance segmentation: Region-based techniques

- Mask Scoring R-CNN (MS-RCNN) [CVPR2019]
  - Examples of the misalignment between mask quality and mask score



Instance segmentation: Region-based techniques

- Mask Scoring R-CNN (MS-RCNN) [CVPR2019]



Instance segmentation: Region-based techniques

- Mask Scoring R-CNN (MS-RCNN) [CVPR2019]
  - http://openaccess.thecvf.com/content_CVPR_2019/papers/Huang_Mask_Scoring_R-CNN_CVPR_2019_paper.pdf
  - Two-Stage instance segmentation

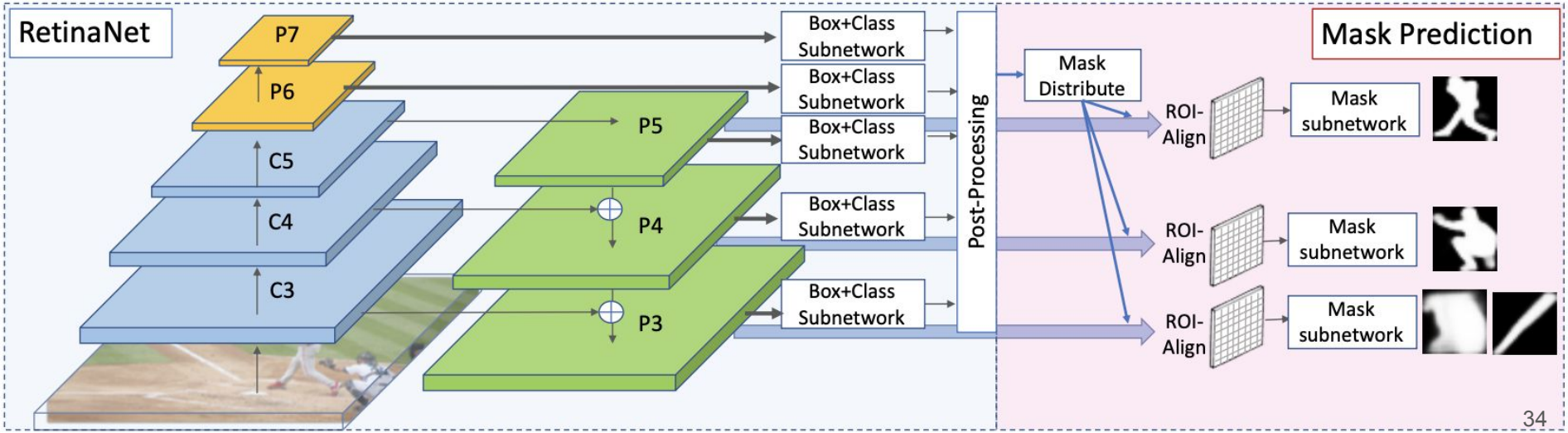
Instance segmentation: Region-based techniques

- **RetinaMask**: Learning to predict masks improves state-of-the-art single-shot detection for free [arxiv2019]
  - **Two-stage detectors** better than single-shot detectors in accuracy-vs-speed trade-off
  - **Single-shot detectors** popular in embedded vision applications (RetinaNet)

  - **Goal**: bring single-shot detectors up to the same level as two-stage detectors
  - Improving RetinaNet (single-shot detector) in three ways:
    - **Integrating instance mask prediction**
    - Making the loss function adaptive and more stable (Uses a Self-Adjusting Smooth L1 Loss)
    - Including hard examples in training
  - **Similar idea from Faster R-CNN -> Mask R-CNN**

Instance segmentation: Region-based techniques

- **RetinaMask**: Learning to predict masks improves state-of-the-art single-shot detection for free [arxiv2019]
  - P5 predicts masks for larger objects
  - P3 predicts masks for smaller objects

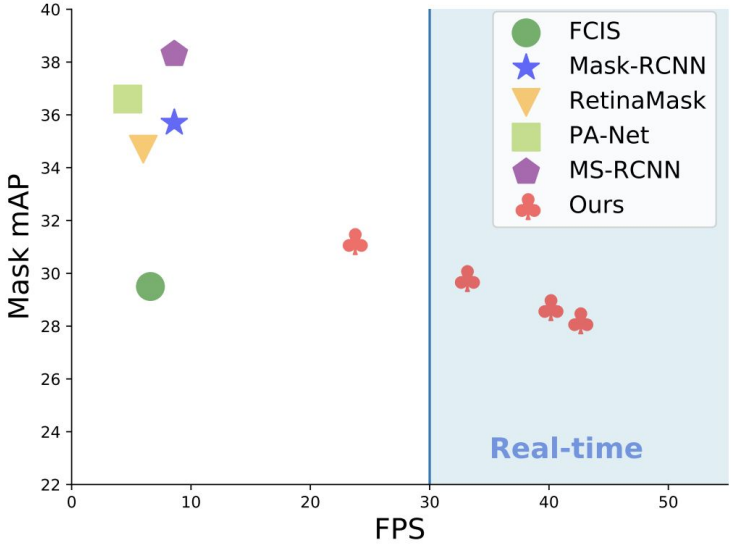


Instance segmentation: Region-based techniques

- **RetinaMask**: Learning to predict masks improves state-of-the-art single-shot detection for free [arxiv2019]
  - https://arxiv.org/pdf/1901.03353.pdf
  - Single-Stage instance segmentation
- **Note that:**
  - Single-stage are usually faster and simpler than two-stage methods.
  - Single-stage methods tend to be less accurate than two-stage methods

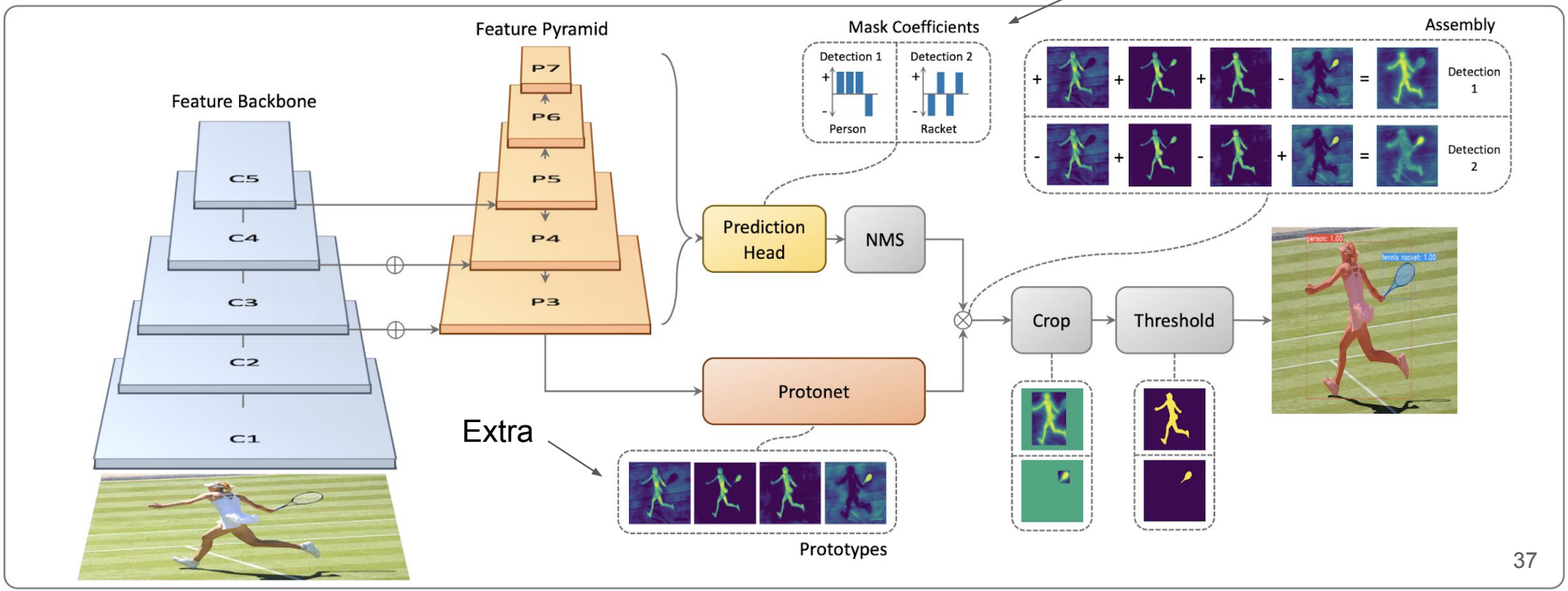
Instance segmentation: Region-based techniques

- **YOLACT** Real-time Instance Segmentation [ICCV2019]
  - Real-time instance segmentation
  - Instance segmentation is broken into two tasks:
    - **Task1**: Generating a set of prototype masks
    - **Task2**: Predicting per-instance mask coefficients
  - Instance masks are produced by linearly combining the prototypes with the mask coefficients



Instance segmentation: Region-based techniques

- YOLACT Real-time Instance Segmentation [ICCV2019]

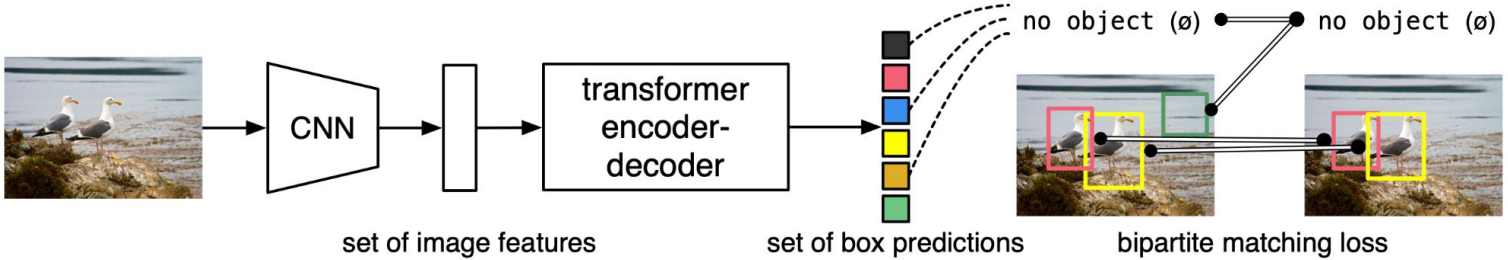


Instance segmentation: Region-based techniques

- YOLACT Real-time Instance Segmentation [ICCV2019]
  - http://openaccess.thecvf.com/content_ICCV_2019/papers/Bolya_YOLACT_Real-Time_Instance_Segmentation_ICCV_2019_paper.pdf
  - Single-Stage instance segmentation

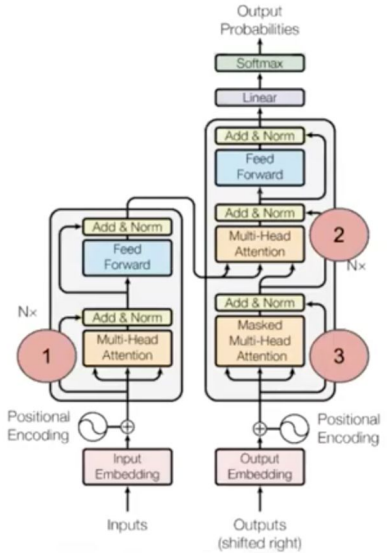
Instance segmentation: Region-based techniques

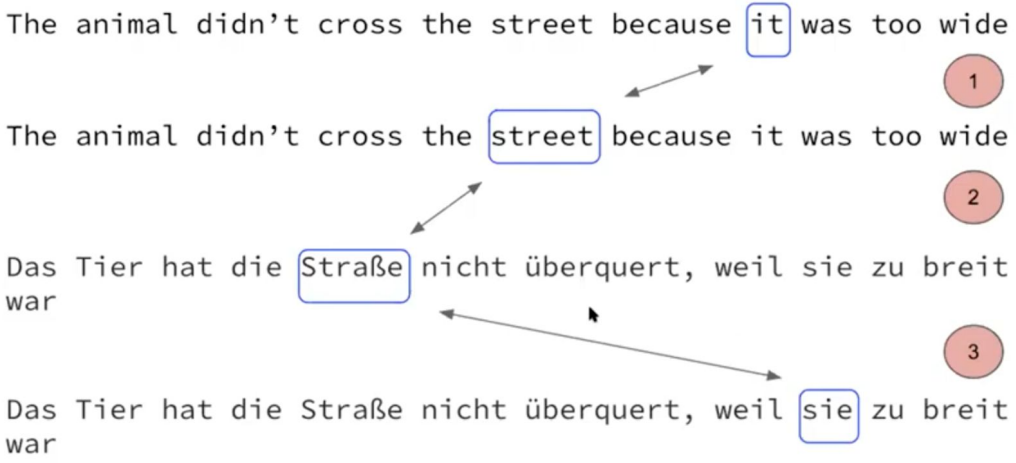
- End-to-end Object Detection with Transformers (DETR)
  - A transformer-based instance segmentation method
    - uses a single neural network to predict object bounding boxes and class labels directly
    - no need for region proposal networks
  - It consists of a **set-based global loss**, makes predictions via
    - bipartite matching to assign predicted boxes to ground-truth objects
  - Given a fixed small set of learned object queries, DETR reasons about the relations of the objects and the global image context to directly output the final set of predictions in parallel.
  - Due to this parallel nature, DETR is very fast and efficient.



Instance segmentation: Region-based techniques

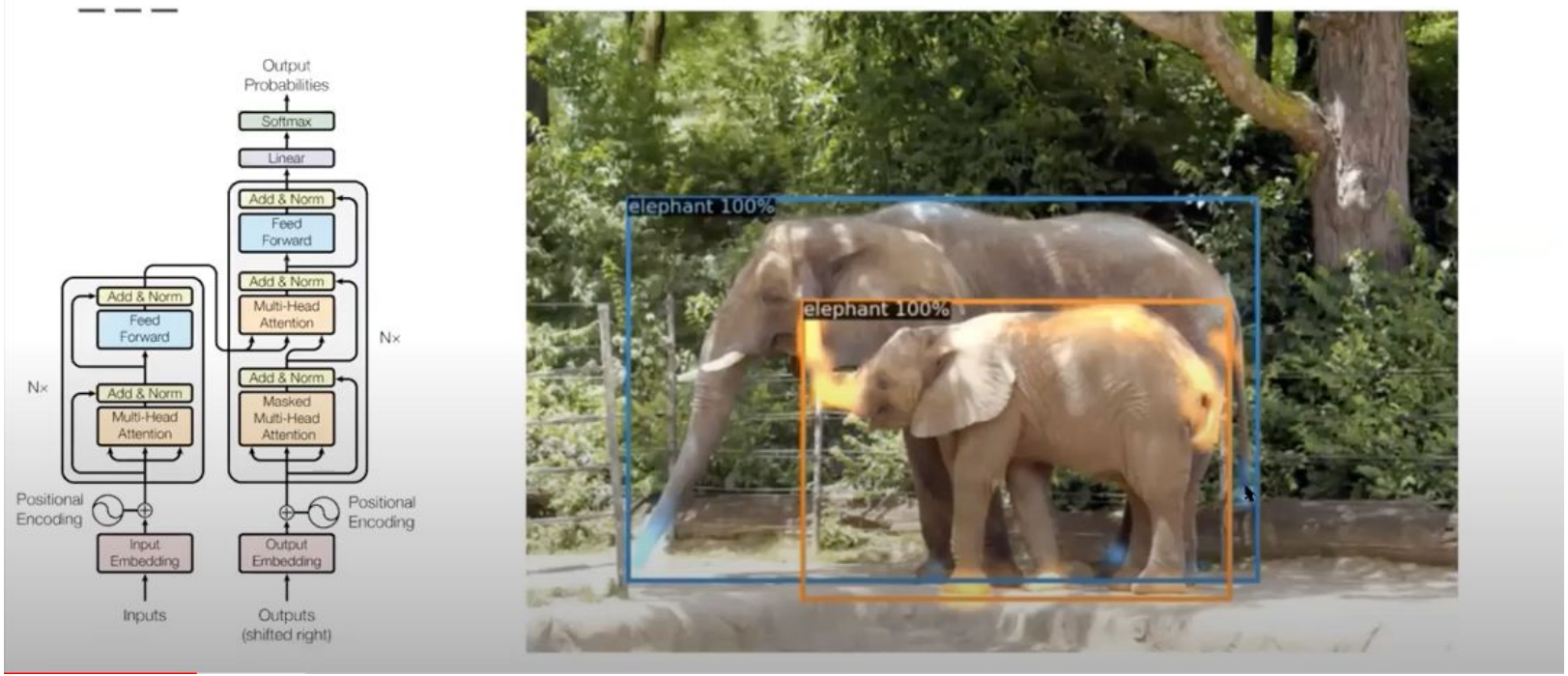
- End-to-end Object Detection with Transformers (DETR)





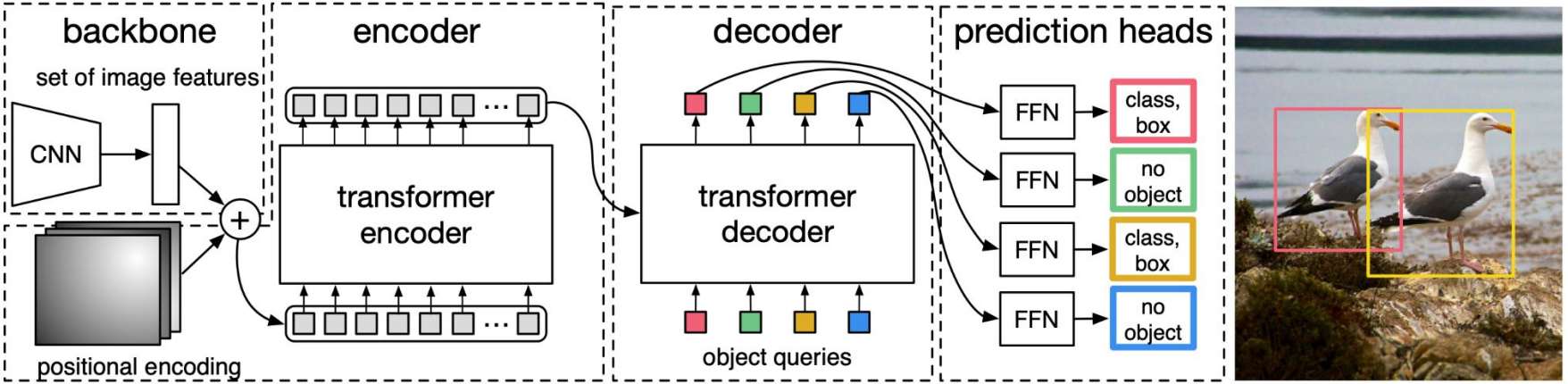
Instance segmentation: Region-based techniques

- End-to-end Object Detection with Transformers (DETR)



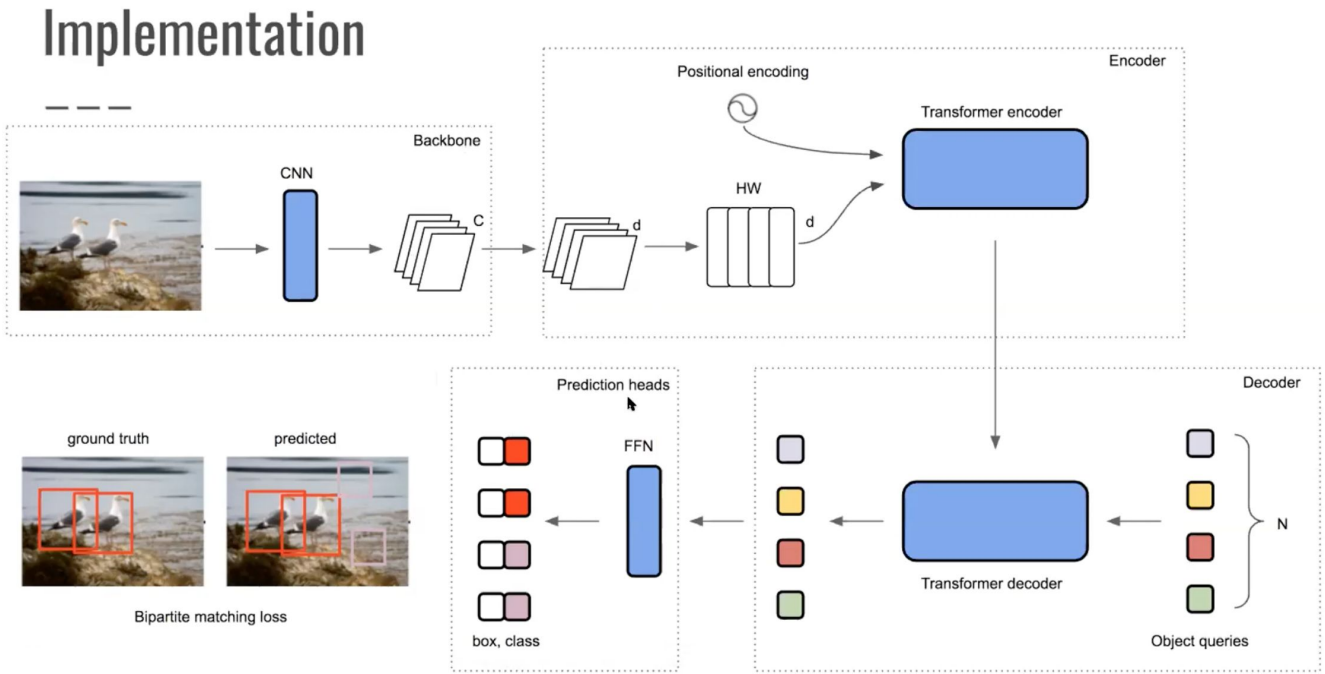
Instance segmentation: Region-based techniques

- End-to-end Object Detection with Transformers (DETR)



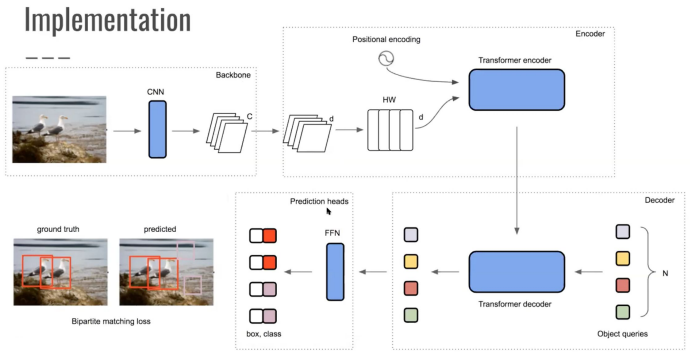
Instance segmentation: Region-based techniques

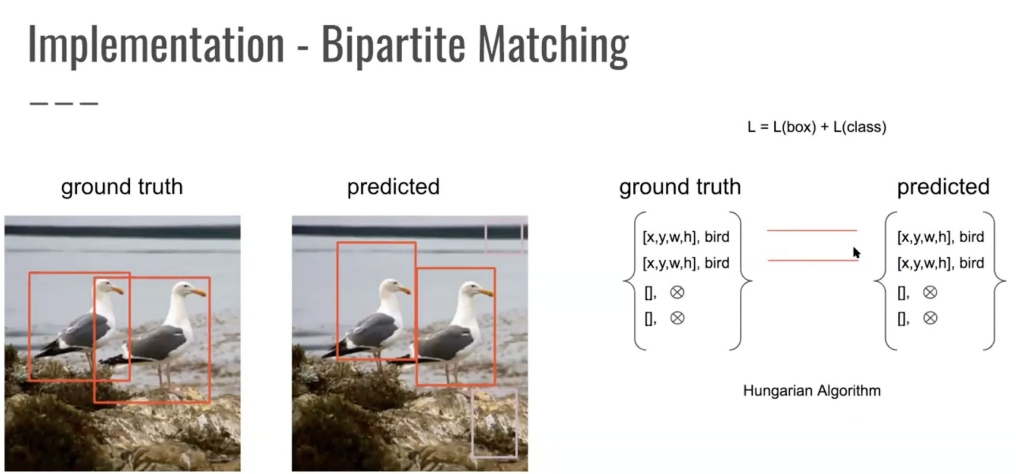
- End-to-end Object Detection with Transformers (DETR)



Instance segmentation: Region-based techniques

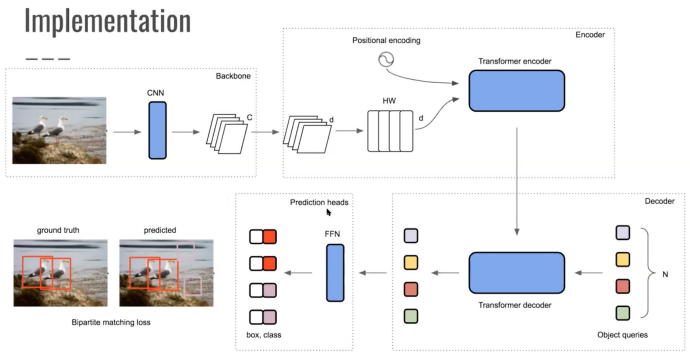
- End-to-end Object Detection with Transformers (DETR)

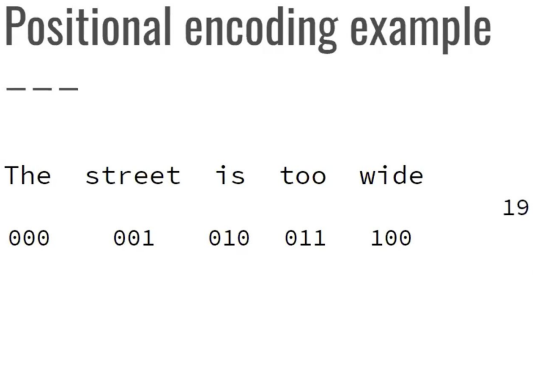




Instance segmentation: Region-based techniques

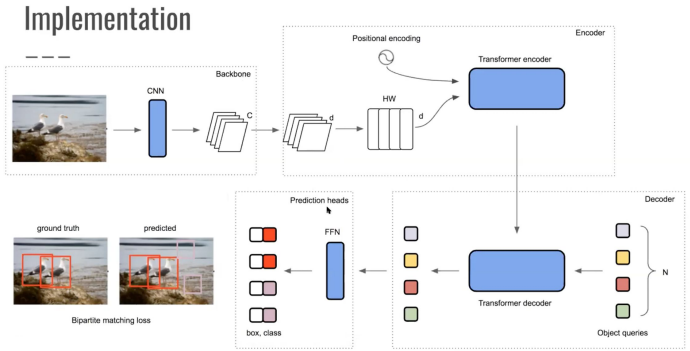
- End-to-end Object Detection with Transformers (DETR)

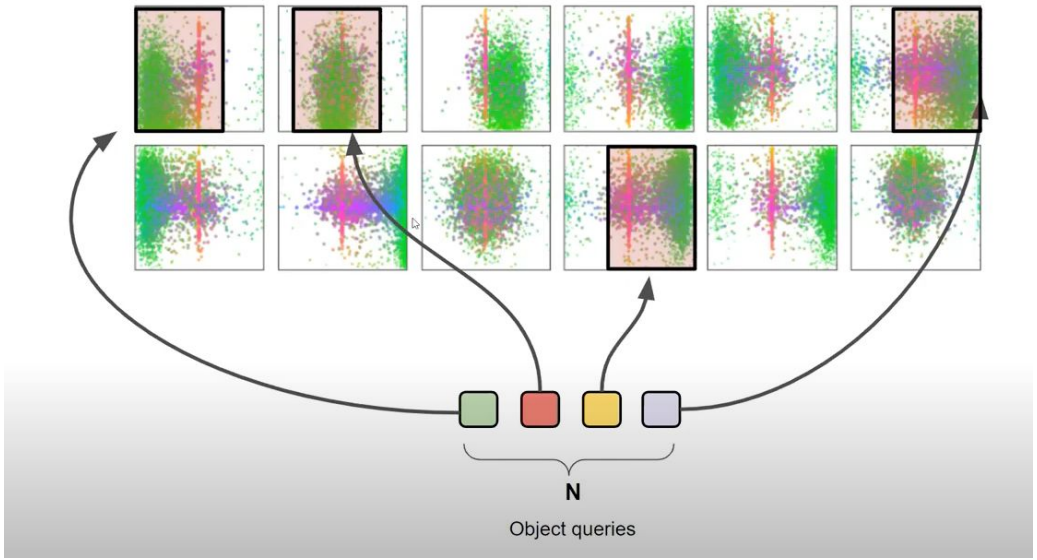




Instance segmentation: Region-based techniques

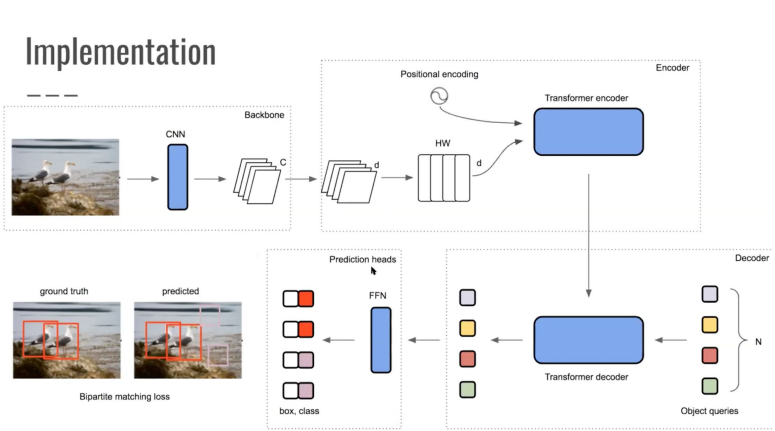
- End-to-end Object Detection with Transformers (DETR)

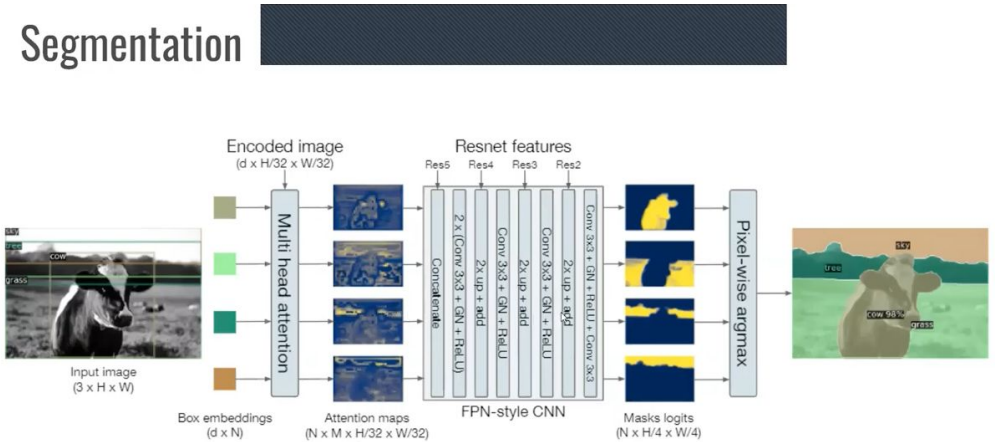




Instance segmentation: Region-based techniques

- End-to-end Object Detection with Transformers (DETR)





Instance segmentation: Region-based techniques

- End-to-end Object Detection with Transformers (DETR)
  - Link

Instance segmentation: RNN-based techniques

- Recurrent Instance Segmentation [ECCV2016] (~200 citations)

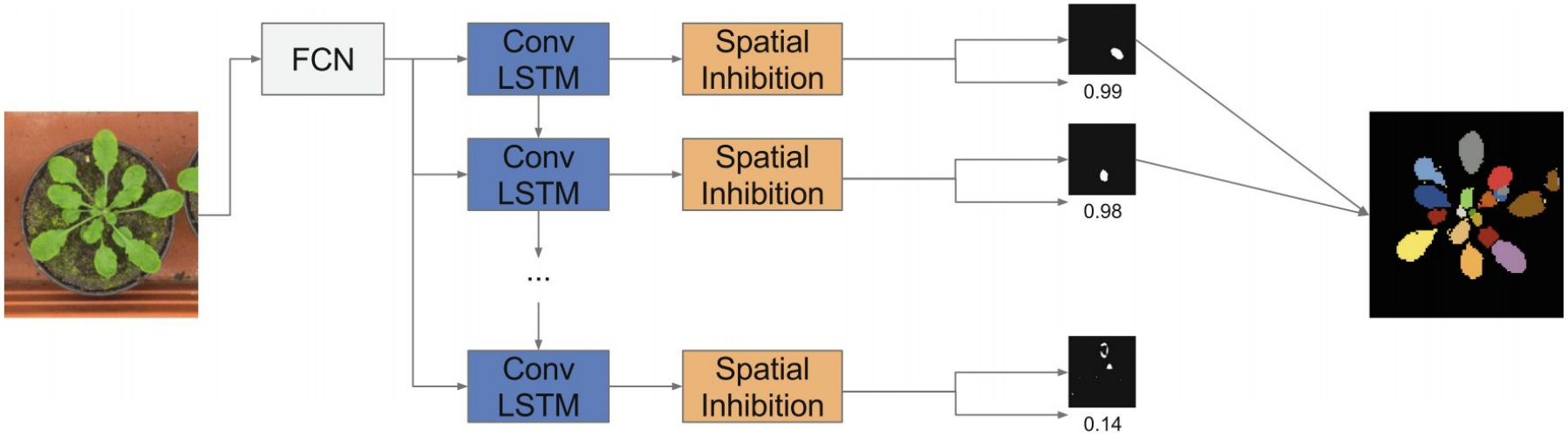


Fig. 1. Diagram of Recurrent Instance Segmentation (RIS).

Instance segmentation: RNN-based techniques

- Recurrent Instance Segmentation [ECCV2016] (~200 citations)
  - Represent instances as a sequence
  - An RNN predicts one instance at a time
  - A spatial memory holds the already segmented pixels
  - Attention via spatial inhibition
  - Mixture of CNN with LSTM (ConvLSTM)

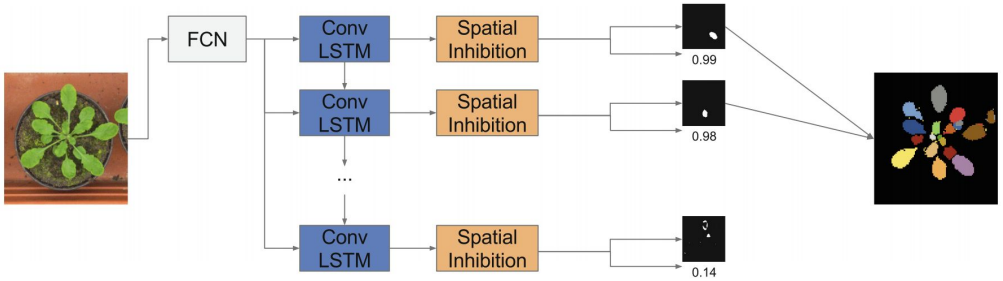


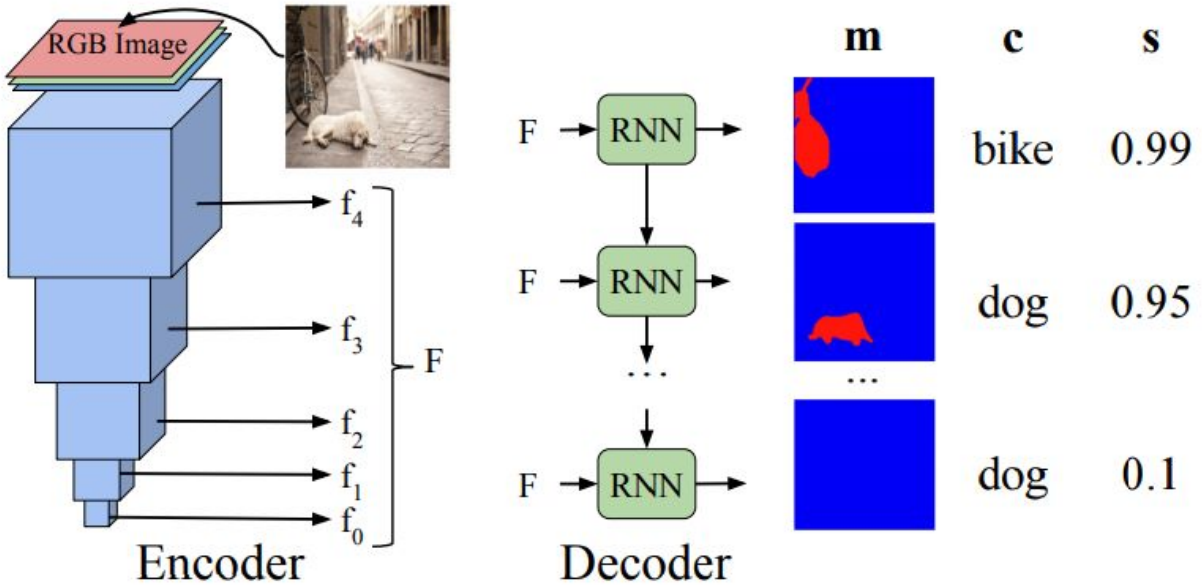
Fig. 1. Diagram of Recurrent Instance Segmentation (RIS).

Instance segmentation: RNN-based techniques

- Recurrent Instance Segmentation [ECCV2016] (~200 citations)
  - arxiv version: https://arxiv.org/pdf/1511.08250.pdf

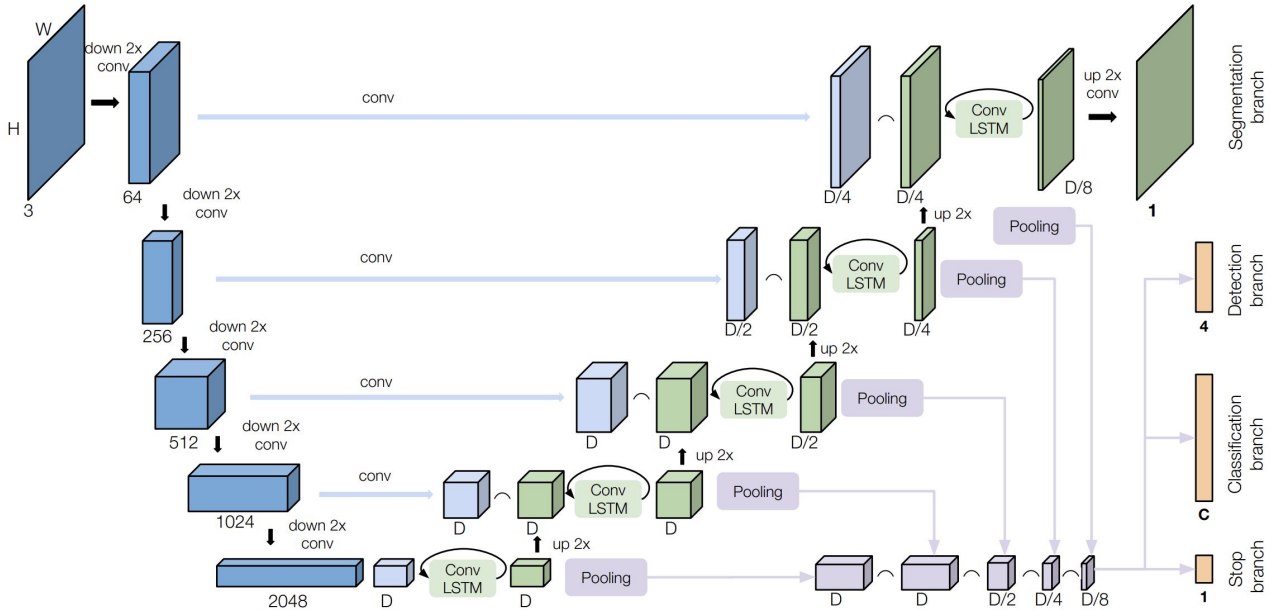
Instance segmentation: RNN-based techniques

- Recurrent Neural Network for Semantic Instance Segmentation (RSIS) [arxiv2017] (~10 citations)



Instance segmentation: RNN-based techniques

- Recurrent Neural Network for Semantic Instance Segmentation (RSIS) [arxiv2017] (~10 citations)



Instance segmentation: RNN-based techniques

- Recurrent Neural Network for Semantic Instance Segmentation (RSIS) [arxiv2017] (~10 citations)

	Rec	Cls	Pascal VOC	CVPPP		Cityscapes			
			$AP_{person,50}$	SBD ↑	DiC ↓	AP	AP_{50}	AP_{car}	$AP_{car,50}$
[17]	✗	✗	–	**84.9**(±**4.8**)	**0.8**(±**1.0**)	**9.5**	**18.9**	**27.5**	41.9
[16]	✓	✗	46.6	56.8(±8.2)	1.1(±0.9)	–	–	–	–
[16] + CRF	✓	✗	50.1	66.6(±8.7)	1.1(±0.9)	–	–	–	–
Ours	✓	✓	**60.7**	74.7(±5.9)	1.1(±0.9)	7.8	17.0	25.8	**45.7**

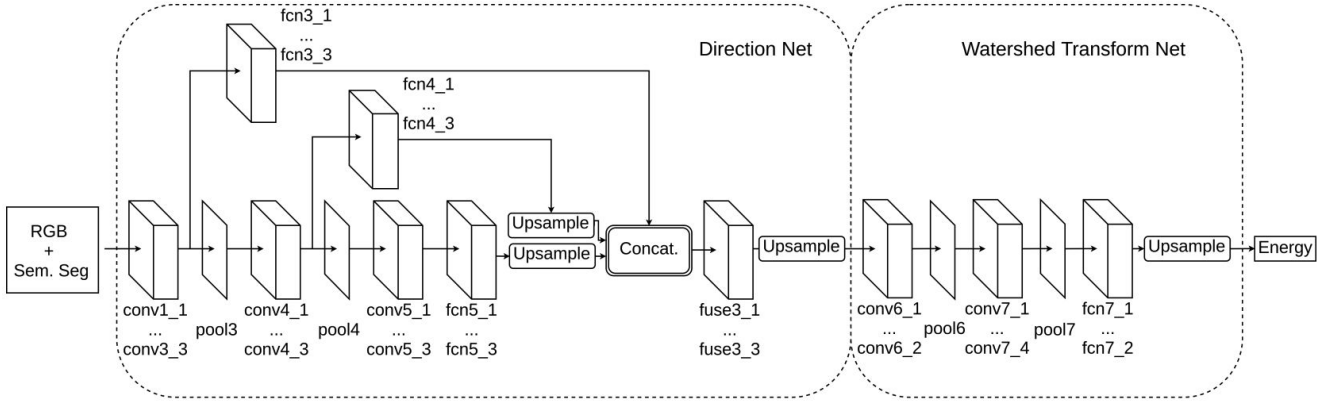
Table 1: Comparison against state of the art sequential methods for semantic instance segmentation. We specify whether the method is recurrent (Rec) and produces categorical probabilities (Cls).

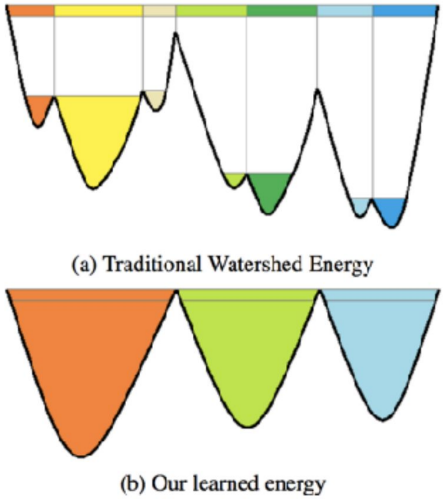
Instance segmentation: RNN-based techniques

- Recurrent Neural Network for Semantic Instance Segmentation (RSIS) [arxiv2017] (~10 citations)
  - arxiv version: https://arxiv.org/pdf/1712.00617.pdf

Instance segmentation: Partition space techniques

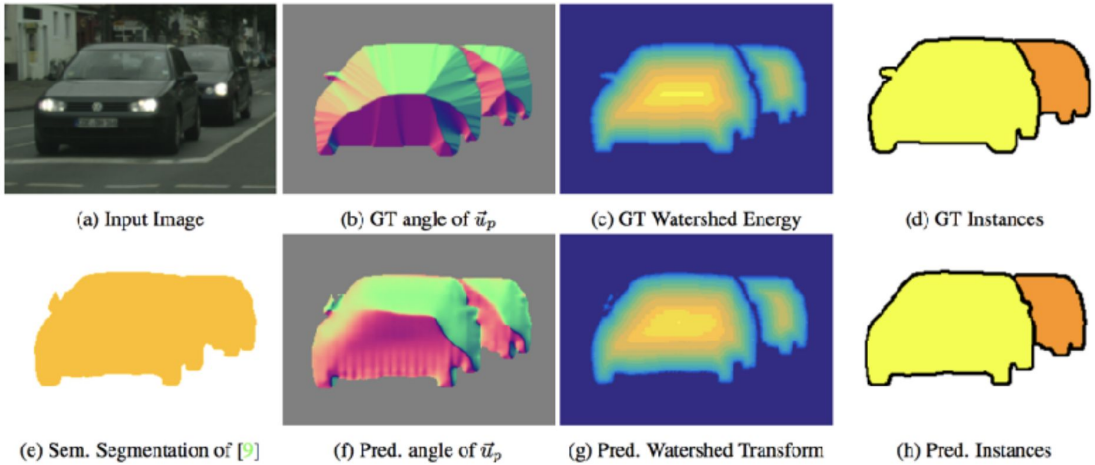
- Deep Watershed Transform for Instance Segmentation [CVPR2017] (~200 citations)
  - Architecture to learn a watershed energy landscape
  - Each **basin** corresponds to an instance
  - **Ridges** are at the same "energy height"
  - Input consists of the **semantic segmentation map**





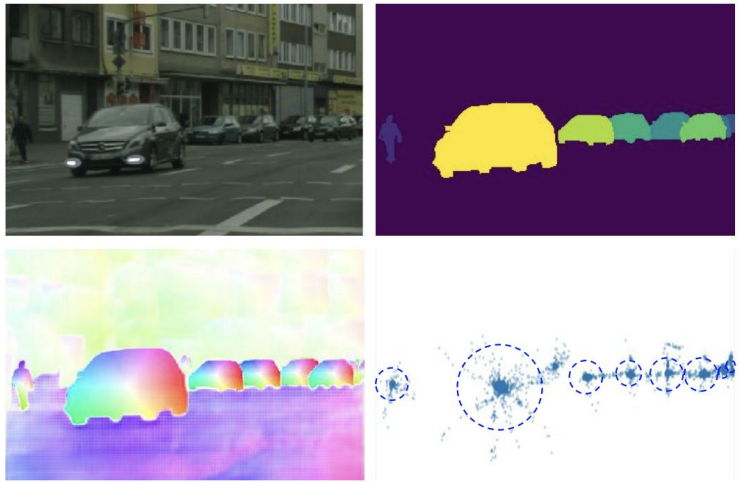
Instance segmentation: Partition space techniques

- Deep Watershed Transform for Instance Segmentation [CVPR2017]
  - Watershed transform as a multi-task learning problem
    - **Task 1:** Learn distance transform of each point to object boundaries
      - Unit vector pointing away from the nearest border pixel
      - Associate wrong object pixel with **maximum angular penalty**
    - **Task 2:** Predict energy function



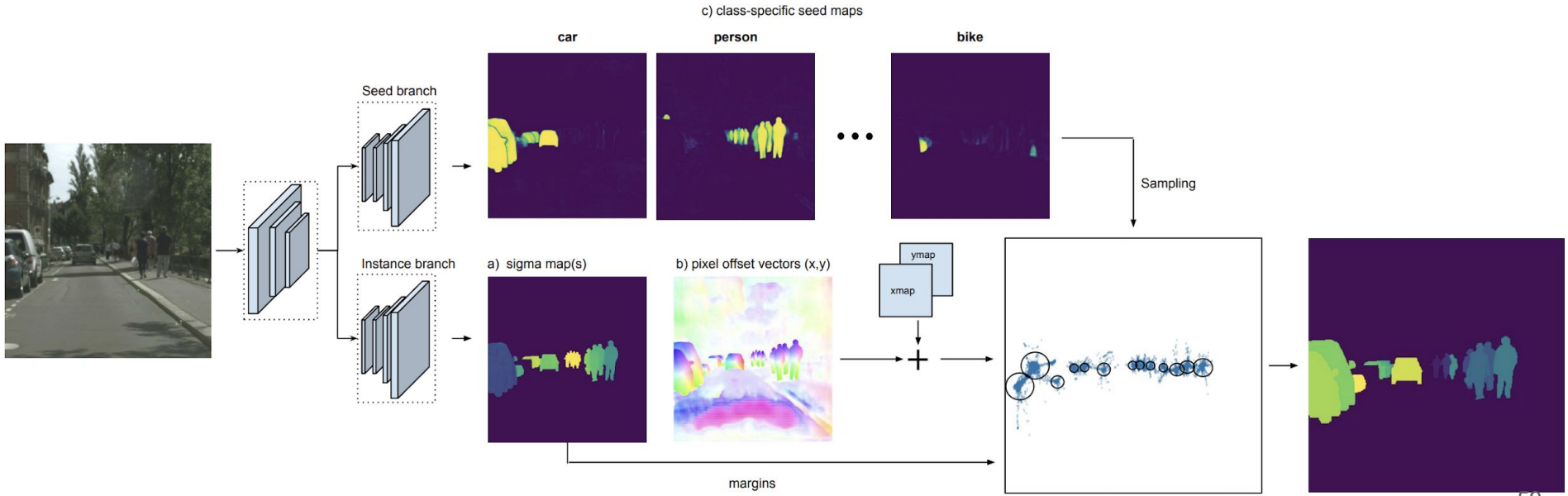
Instance segmentation: Partition space techniques

- Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth [CVPR2019]
  - Proposal-free method for instance segmentation
    - Often faster than proposal-based but with lower accuracy
  - Clustering loss
    - It pulls the spatial embeddings of pixels belonging to the same instance together
  - Real-time with high accuracy



Instance segmentation: Partition space techniques

- Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth [CVPR2019]



Instance segmentation: Partition space techniques

- Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth [CVPR2019]
  - http://openaccess.thecvf.com/content_CVPR_2019/papers/Neven_Instance_Segmentation_by_Jointly_Optimizing_Spatial_Embeddings_and_Clustering_Bandwidth_CVPR_2019_paper.pdf

Instance segmentation: datasets

- Many datasets for semantic segmentation are also used for instance segmentation
  - **MS COCO -> most commonly used**
  - Cityscapes
  - Mapillary Vistas

Instance segmentation: datasets

- New datasets have appeared:
  - **Open Images V6 (Feb 2020)**
    - ~9M images annotated with image-level labels, object bounding boxes, object segmentation masks and visual relationships
    - 8.3 objects per image on average
    - **2.8M object instances annotated with segmentation masks in 350 classes**
      - MS COCO (1.5M object instances in 80 classes)

Instance segmentation: datasets

- New datasets have appeared:
  - Open Images V6 (Feb 2020)

Instance segmentation: datasets

- New datasets have appeared:
  - **LVIS**: A Dataset for Large Vocabulary Instance Segmentation [CVPR2019]
    - 164K images (less than MS COCO and Open Images)
    - 1000 object categories (vs 350 in Open Images)
    - 2.2M high-quality instance masks (similar to MS COCO and Open Images)
    - 11.2 objects instance from 3.4 categories on average per image (more complex images than Open Images and MS COCO)
    - Useful for few-shot object detection

Instance segmentation: datasets

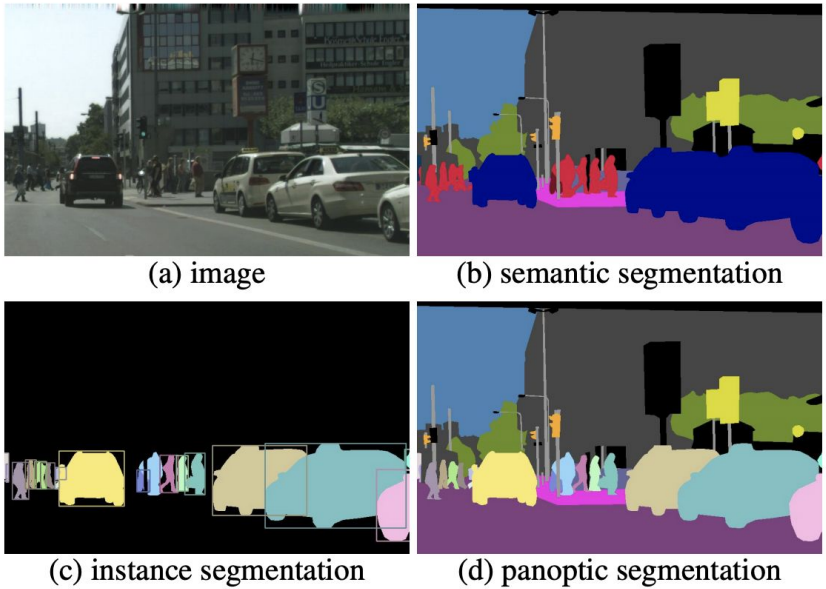
- New datasets have appeared:
  - **LVIS**: A Dataset for Large Vocabulary Instance Segmentation [CVPR2019]

Outline

- Introduction to segmentation
- Semantic segmentation
- Instance segmentation
- **Panoptic segmentation**
- Amodal segmentation
- Referring image segmentation
- Current trends and future research

Panoptic segmentation: problem statement

- It unifies two distinct tasks:
  - Semantic segmentation (assign a class label to each pixel ~ **stuff**)
  - Instance segmentation (detect and segment each object instance ~ **things**)



Panoptic segmentation: problem statement

- New metric for evaluation: panoptic quality (PQ)
  - Captures performance for all classes (things and stuff)
  - Insensitive to class imbalance

$$\mathrm{PQ} = \underbrace{\frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}}$$

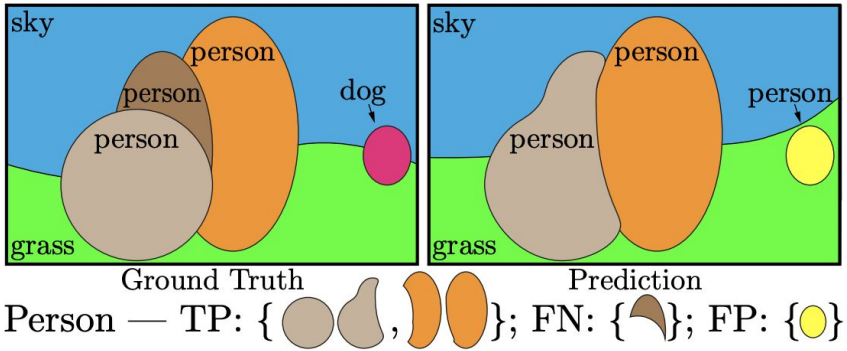
Panoptic segmentation: problem statement

- New metric for evaluation: panoptic quality (PQ)
  - Captures performance for all classes (things and stuff)
  - Insensitive to class imbalance

$$\mathrm{PQ} = \frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$

average IoU of matched segments

penalization of segments without matching



Panoptic segmentation: problem statement

- New task proposed in CVPR2019
  - http://openaccess.thecvf.com/content_CVPR_2019/papers/Kirillov_Panoptic_Segmentation_CVPR_2019_paper.pdf

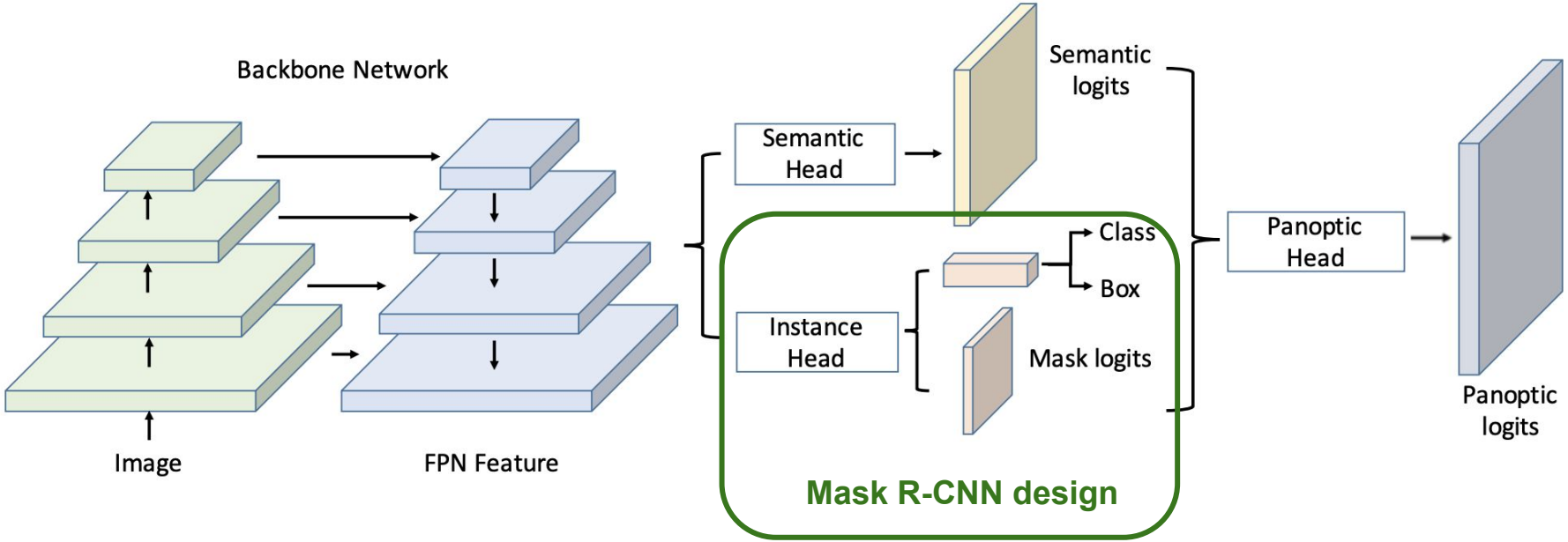
Panoptic segmentation: datasets

- MS COCO
- Cityscapes
- ADE20k
- Mapillary Vistas

COCO 2019 Panoptic Segmentation Task

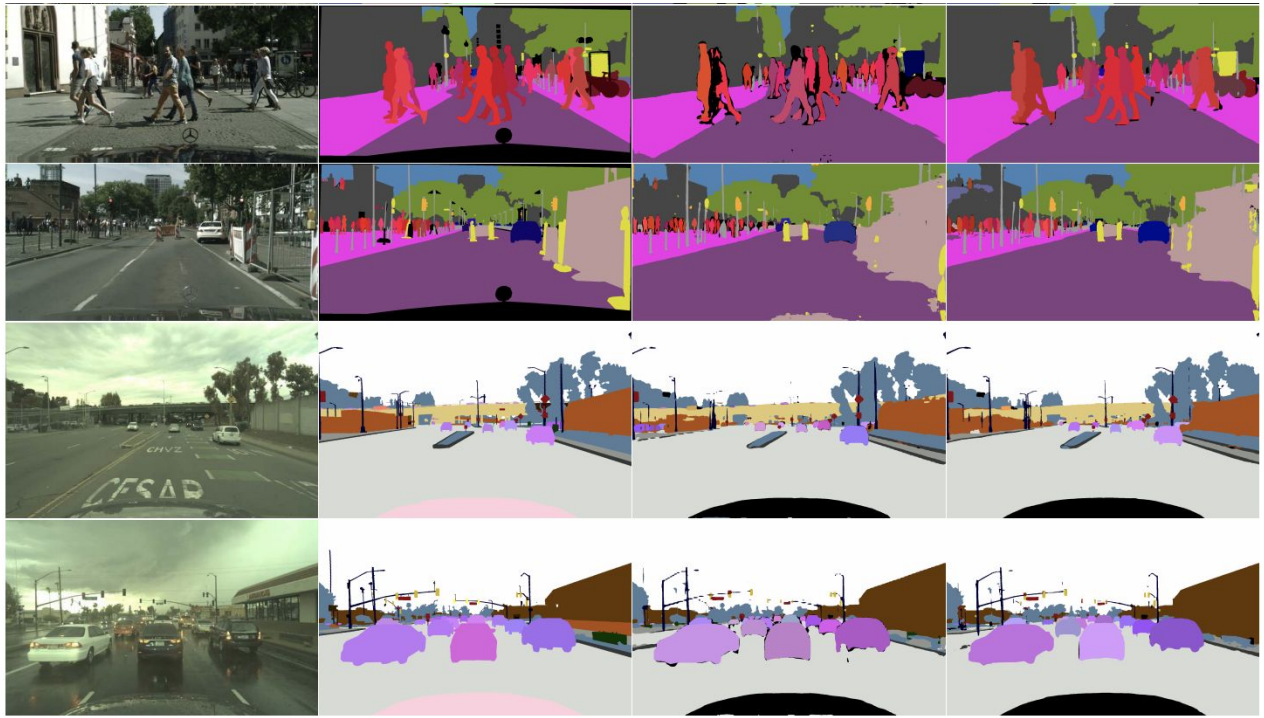
Panoptic segmentation: techniques

- UPSNet: A Unified Panoptic Segmentation Network [CVPR2019]



Panoptic segmentation: techniques

- UPSNet: A Unified Panoptic Segmentation Network [CVPR2019]



Panoptic segmentation: techniques

- UPSNet: A Unified Panoptic Segmentation Network [CVPR2019]
  - http://openaccess.thecvf.com/content_CVPR_2019/papers/Xiong_UPSNet_A_Unified_Panoptic_Segmentation_Network_CVPR_2019_paper.pdf

Outline

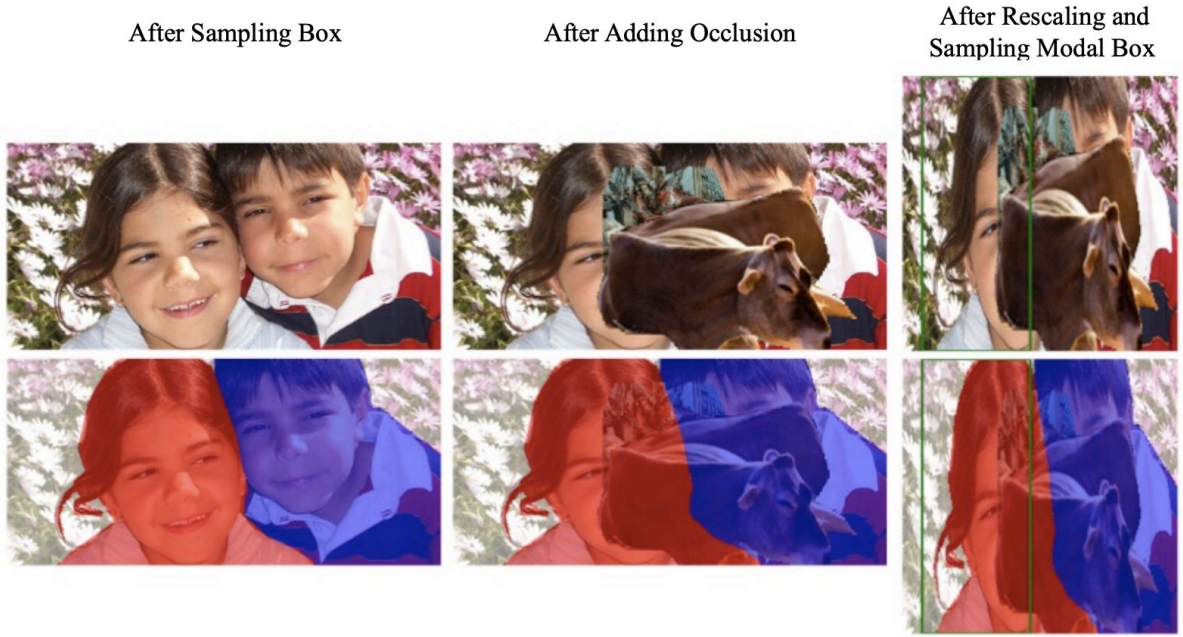
- Introduction to segmentation
- Semantic segmentation
- Instance segmentation
- Panoptic segmentation
- **Amodal segmentation**
- Referring image segmentation
- Current trends and future research

Amodal segmentation: problem definition

- Objective: predict the region encompassing both **visible and occluded** parts of each object.
- Problem defined in [ECCV2016]

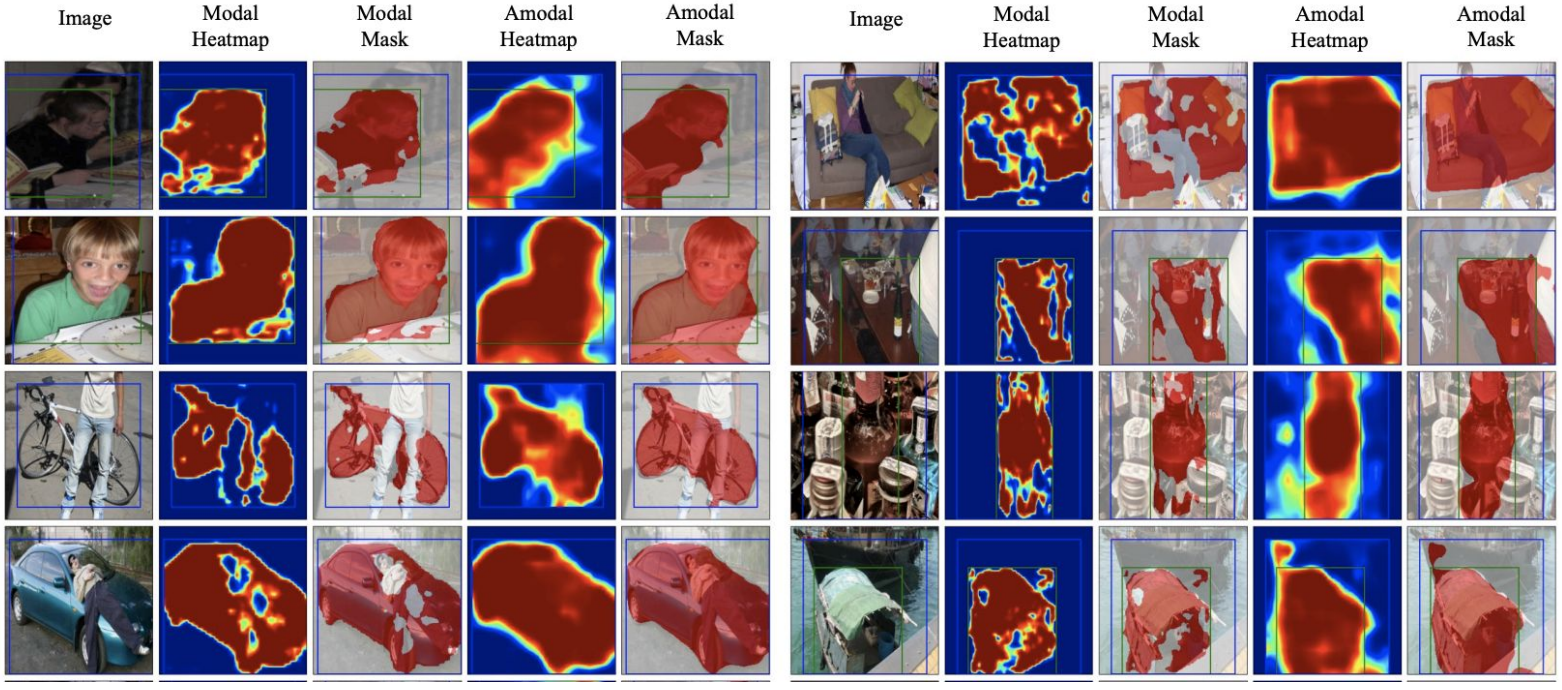
Amodal segmentation: techniques

- Amodal instance segmentation [ECCV2016]
  - Training data from modal segmentation problem
    - **Adding occlusions**



Amodal segmentation: techniques

- Amodal instance segmentation [ECCV2016]
  - Visual results



Amodal segmentation: techniques

- Amodal instance segmentation [ECCV2016]
  - https://arxiv.org/pdf/1604.08202.pdf

Amodal segmentation: techniques

- Semantic Amodal Segmentation [CVPR2017]
  - A detailed image annotation that captures information beyond the visible pixels and requires **complex reasoning about full scene structure**.
  - An amodal segmentation of each image is created:
    - the full extent of each region is marked, **not just the visible pixels**.
  - **Two datasets** for semantic amodal segmentation are created:
    - 500 images from BSDS dataset
    - 5000 images from COCO

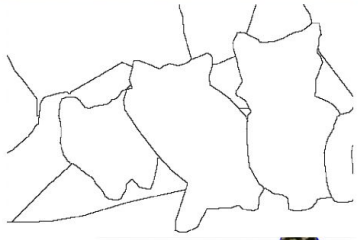
Amodal segmentation: techniques

- Semantic Amodal Segmentation [CVPR2017]

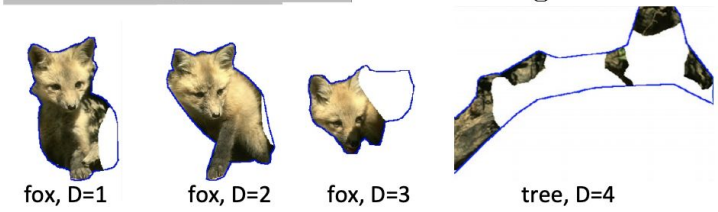
input image

modal segmentation

depth ordering



edge annotations



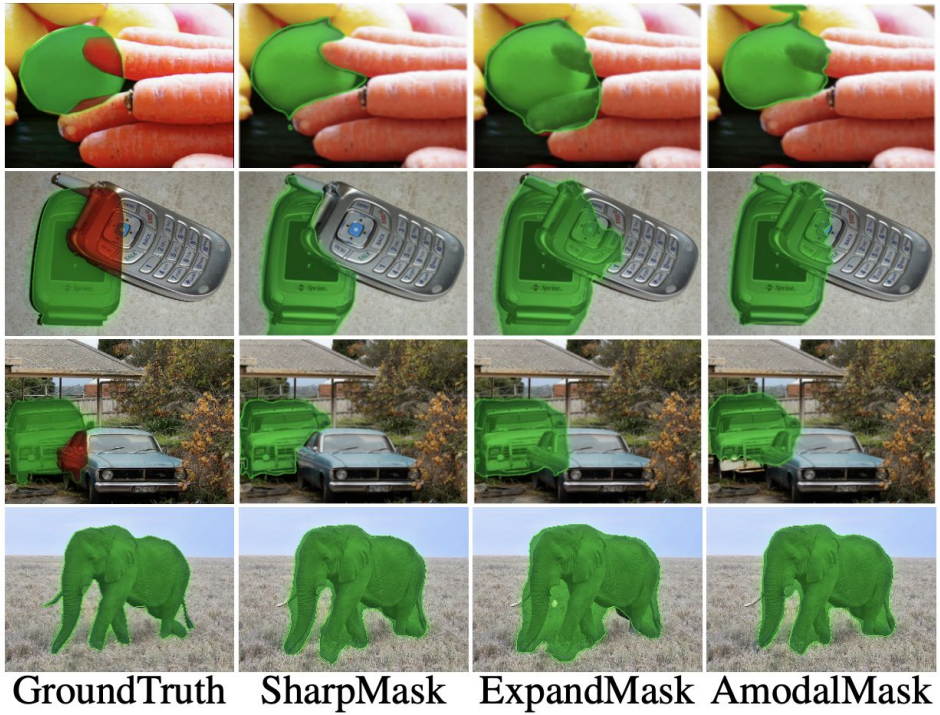
amodal segmentation
+
semantic + depth order

Amodal segmentation: techniques

- Semantic Amodal Segmentation [CVPR2017]
  - **Baselines**:
    - **ExpandMask**: network that takes an image patch and a modal mask generated by SharpMask as input and outputs an amodal mask
    - **AmodalMask**: network that directly predict amodal masks from image patches
  - **Metrics**:
    - Average Recall (AR) at multiple IoU thresholds
      - Same metric as instance segmentation but uses IoU against amodal masks

Amodal segmentation: techniques

- Semantic Amodal Segmentation [CVPR2017]
  - **Visual results**



Amodal segmentation: techniques

- Semantic Amodal Segmentation [CVPR2017]
  - http://openaccess.thecvf.com/content_cvpr_2017/papers/Zhu_Semantic_Amodal_Segmentation_CVPR_2017_paper.pdf

Outline

- Introduction to segmentation
- Semantic segmentation
- Instance segmentation
- Panoptic segmentation
- Amodal segmentation
- **Referring image segmentation**
- Current trends and future research

Referring image segmentation: problem definition

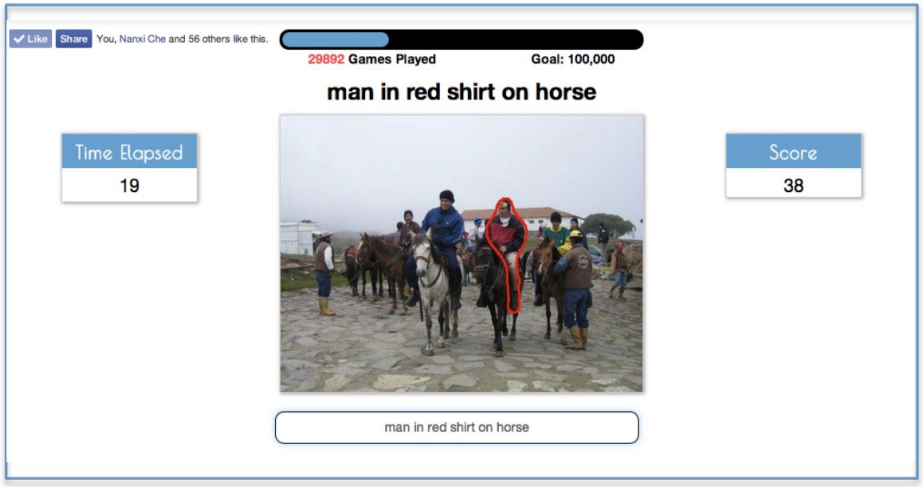
- The task of referring expression comprehension is to localize a region described by a given referring expression

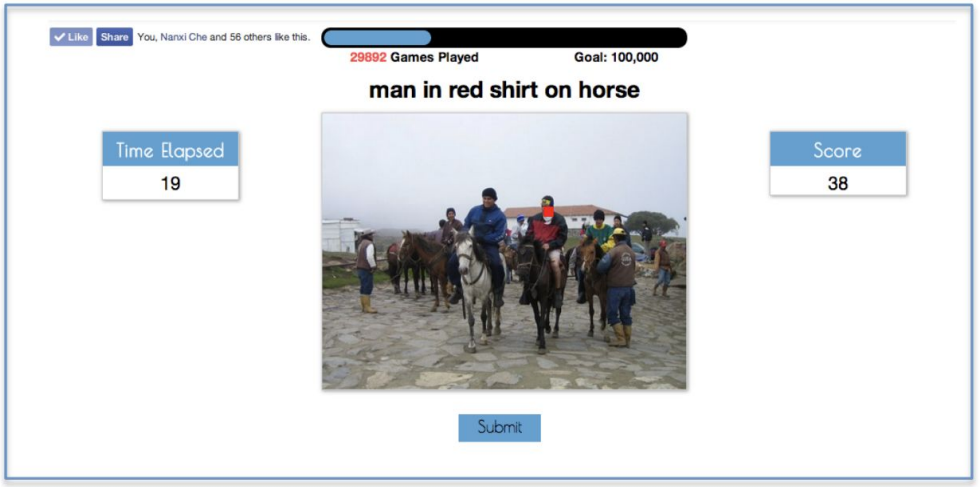
Referring image segmentation: problem definition

- Assumption about referring expressions:
  - A referring expression can't be ambiguous, it should identify the object from the scene without any ambiguity
- Modeling Context in Referring Expressions[ECCV2016]
- 4 principles about cooperative natural language dialogue interactions:
  - Quality (try to be truthful)
  - Quantity (make your contribution as informative as you can, giving as much information as is needed but no more)
  - Relevance (be relevant and pertinent to the discussion)
  - Manner (be as clear, brief, and orderly as possible, avoiding obscurity and ambiguity).

Referring image segmentation: problem definition

- Referring expressions generation by playing:
  - **ReferItGame**: Referring to Objects in Photographs of Natural Scenes [EMNLP2014]



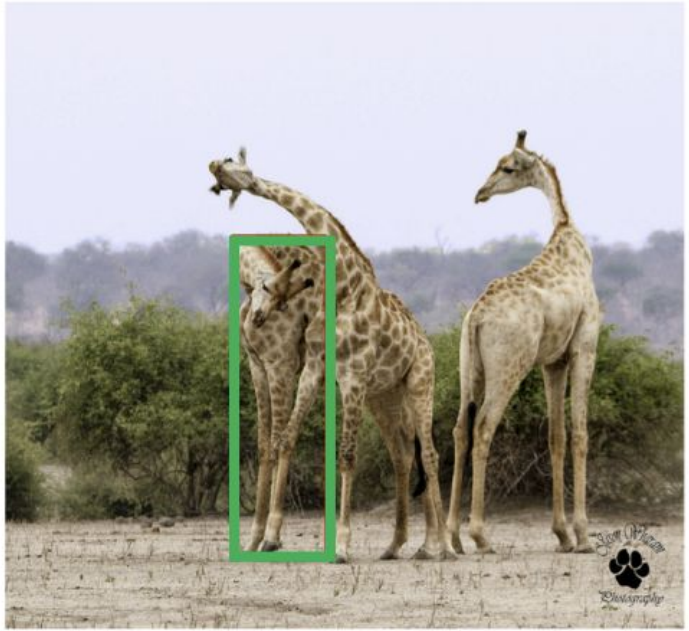


Referring image segmentation: techniques

- Modeling context in referring expressions [ECCV2016]
  - Three datasets based on COCO are proposed:
    - RefCOCOg: referring expressions collected on Amazon Mechanical Turk:
      - one set of workers is asked to write referring expressions for MS COCO images
      - another set of workers is asked to click on the indicated object given a referring expression
    - RefCOCO: referring expressions collected with ReferItGame
      - **No restrictions** are placed on the type of language used
    - RefCOCO+: referring expressions collected with ReferItGame
      - Players are **disallowed from using location words** in their referring expressions
  - RefCOCOg & RefCOCO & RefCOCO+:
    - ~140k referring expressions
    - 50k objects from 20k images

Referring image segmentation: techniques

- Modeling context in referring expressions [ECCV2016]



RefCOCO:
1. giraffe on left
2. first giraffe on left

RefCOCO+:
1. giraffe with lowered head
2. giraffe head down

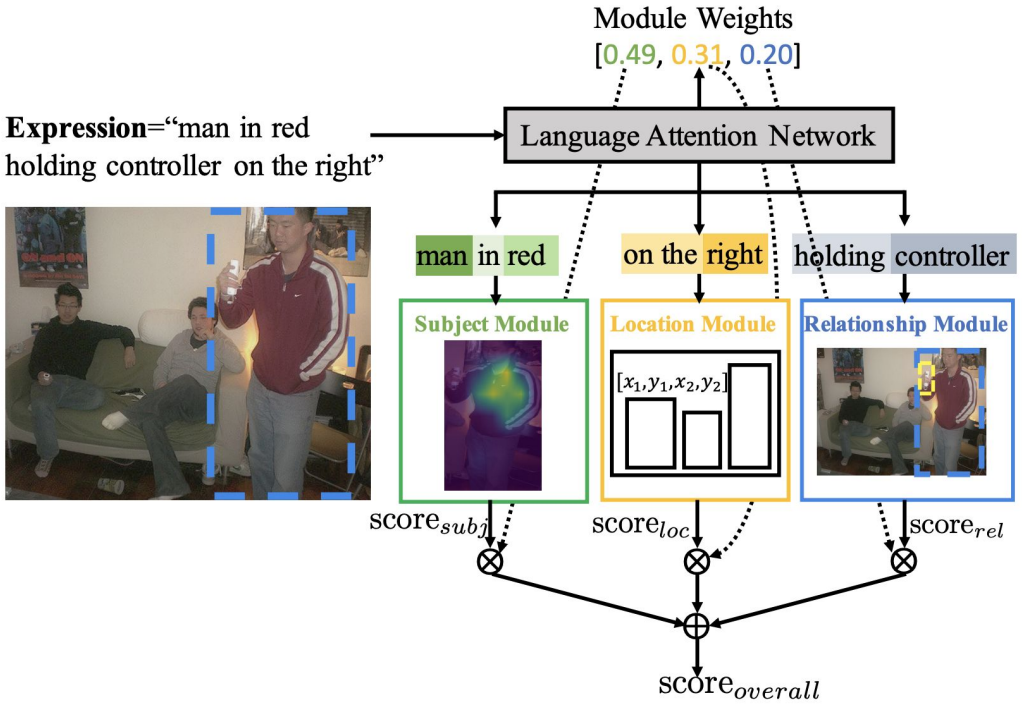
RefCOCOg:
1. an adult giraffe scratching its back with its horn
2. giraffe hugging another giraffe

Referring image segmentation: techniques

- MAttNet: Modular Attention Network for Referring Expression Comprehension [CVPR2018]
  - **Motivation**: most work treat expressions as a **single unit**
  - **Idea**: Decompose the expressions into **three modular components**:
    - appearance
    - location
    - relation to other objects
  - **Two types of attention**:
    - language-based: word/phrase attention that each module should focus on
    - visual attention: relevant image components that each module should focus on

Referring image segmentation: techniques

- MAttNet: Modular Attention Network for Referring Expression Comprehension [

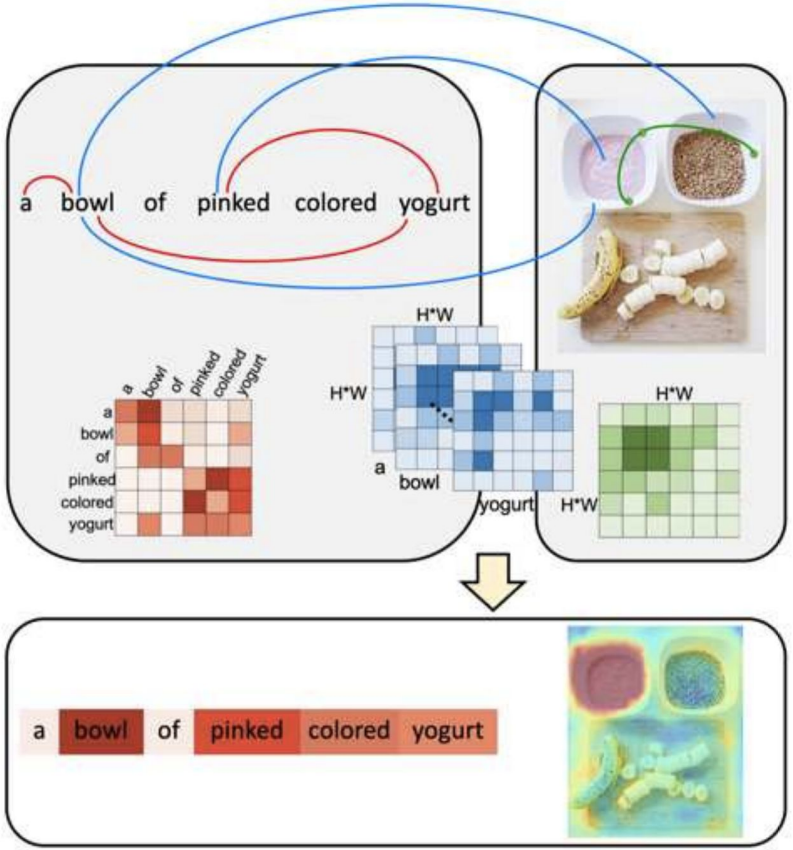


Referring image segmentation: techniques

- Cross-Modal Self-Attention Network for Referring Image Segmentation (CMSA) [CVPR2019]
  - **Motivation**: Existing works treat the language expression and the input image **separately** in their representations.
  - Idea:
    - a **cross-modal** self-attention (CMSA) module that effectively captures the long-range dependencies between linguistic and visual features.
    - adaptively focus on informative words in the referring expression and important regions in the input image

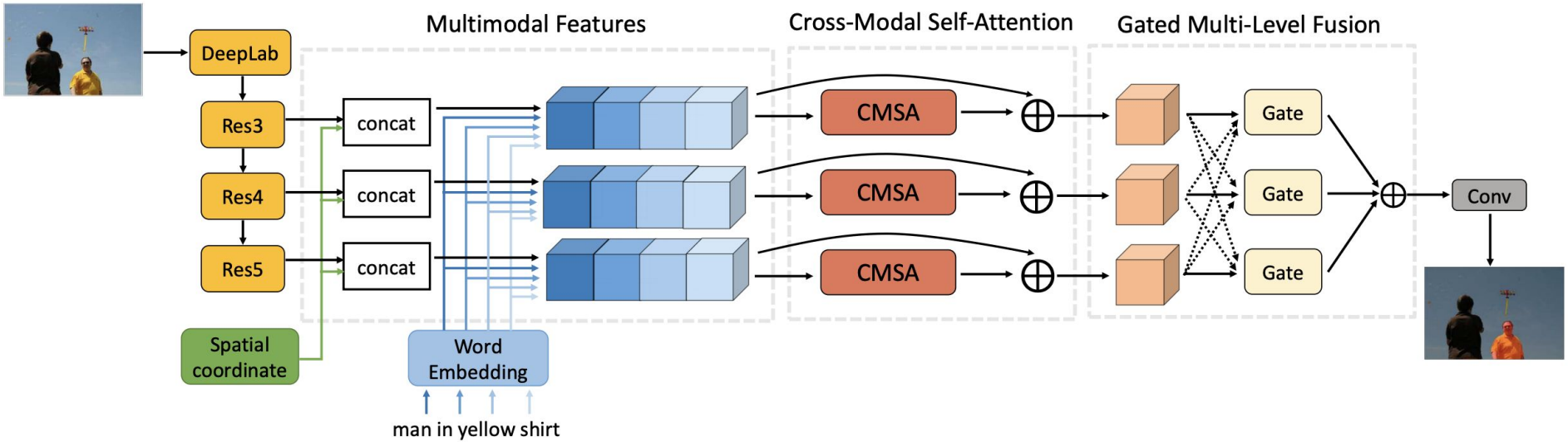
Referring image segmentation: techniques

- Cross-Modal Self-Attention Network for Referring Image Segmentation (CMSA) [CVPR2019]



Referring image segmentation: techniques

- Cross-Modal Self-Attention Network for Referring Image Segmentation (CMSA) [CVPR2019]

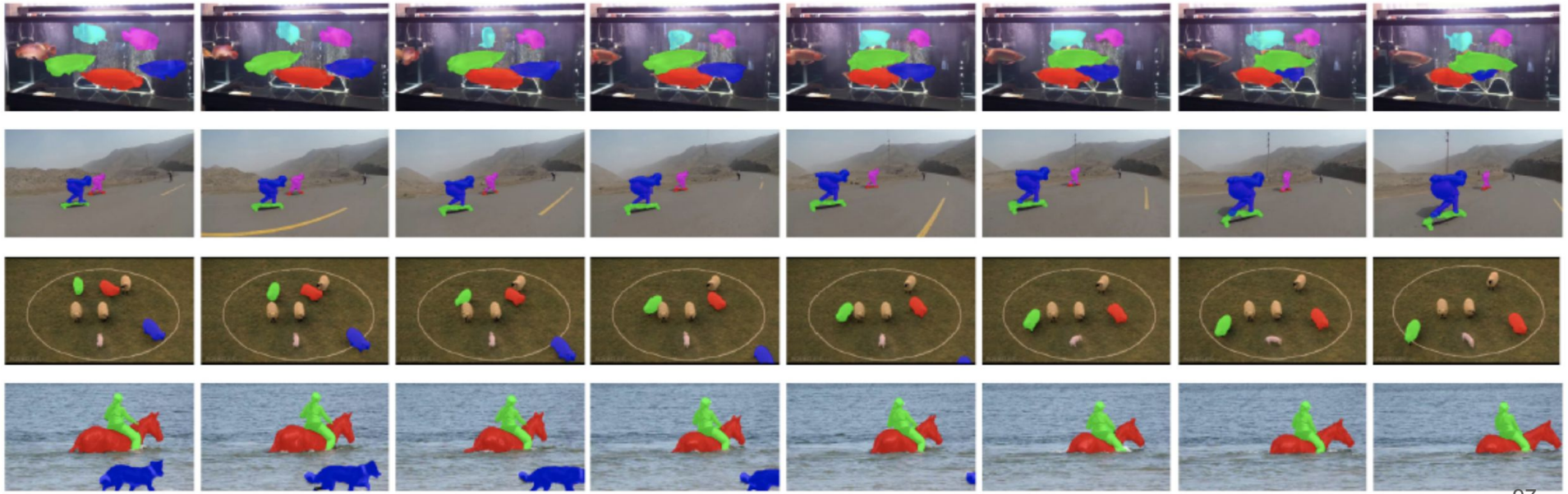


Outline

- Introduction to segmentation
- Semantic segmentation
- Instance segmentation
- Panoptic segmentation
- Amodal segmentation
- Referring image segmentation
- **Current trends and future research**

Current trends and future research

- Video object segmentation
  - RVOS: End-to-End Recurrent Network for Video Object Segmentation [CVPR2019]



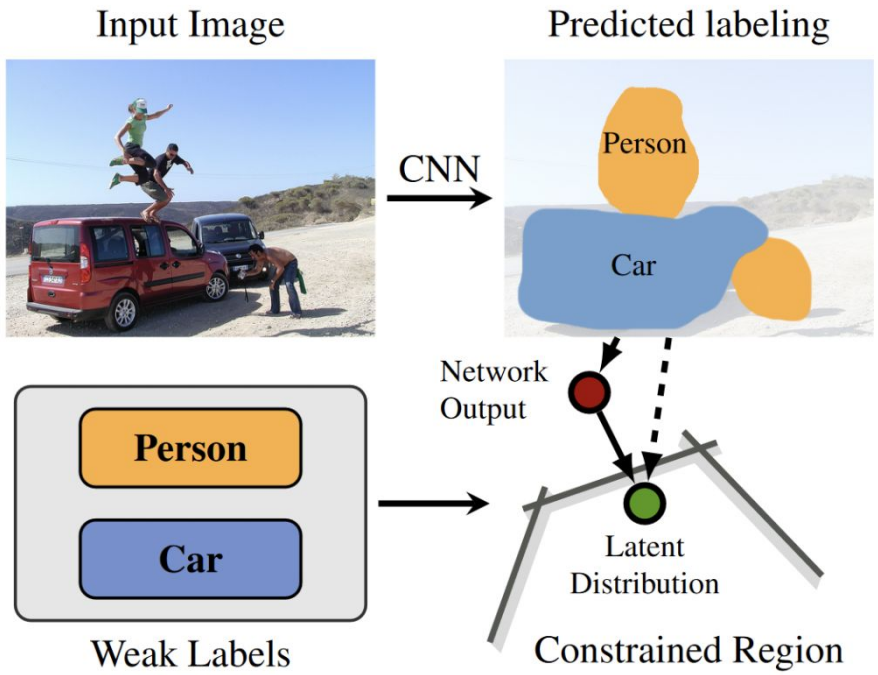
Current trends and future research

- Weakly supervised segmentation
  - Cost of labeling vs. quality of labels
    - Heavy labeling efforts
    - Pixel-wise labeling is expensive
    - Error-prone and hard to be precise
    - Categories can be too numerous…
- Most of my PhD work is in this topic

Image source *https://sthalles.github.io/deep_segmentation_network/*

Current trends and future research

- Weakly supervised segmentation
- How to exploit weaker labels to perform semantic segmentation?
  - Image-level annotations https://arxiv.org/abs/1506.03648

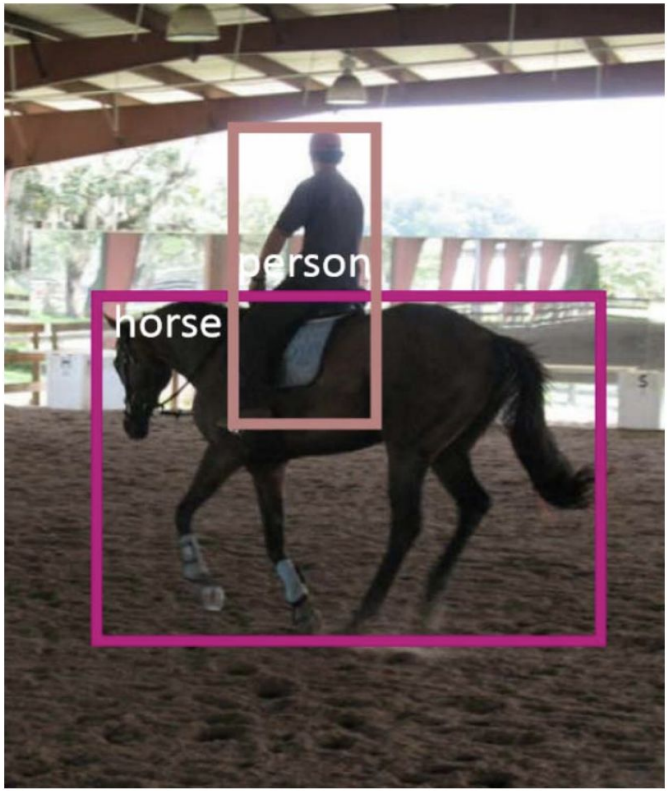


Current trends and future research

- Weakly supervised segmentation
- How to exploit weaker labels to perform semantic segmentation?
  - Image-level annotations https://arxiv.org/abs/1506.03648
  - Point annotations https://arxiv.org/abs/1506.02106

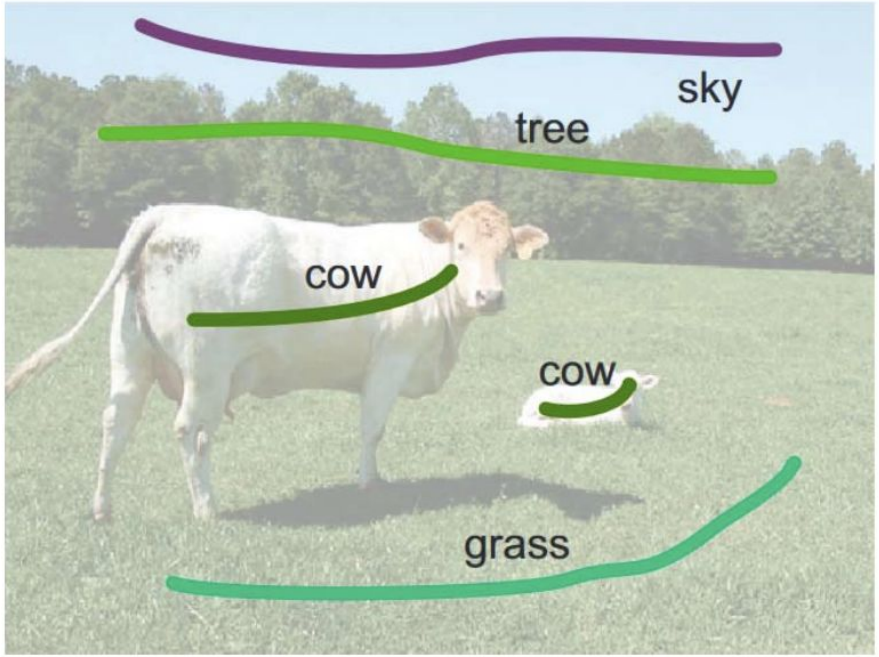
Current trends and future research

- Weakly supervised segmentation
- How to exploit weaker labels to perform semantic segmentation?
  - Image-level annotations
    https://arxiv.org/abs/1506.03648
  - Point annotations
    https://arxiv.org/abs/1506.02106
  - Bounding-box annotations
    https://arxiv.org/abs/1503.01640



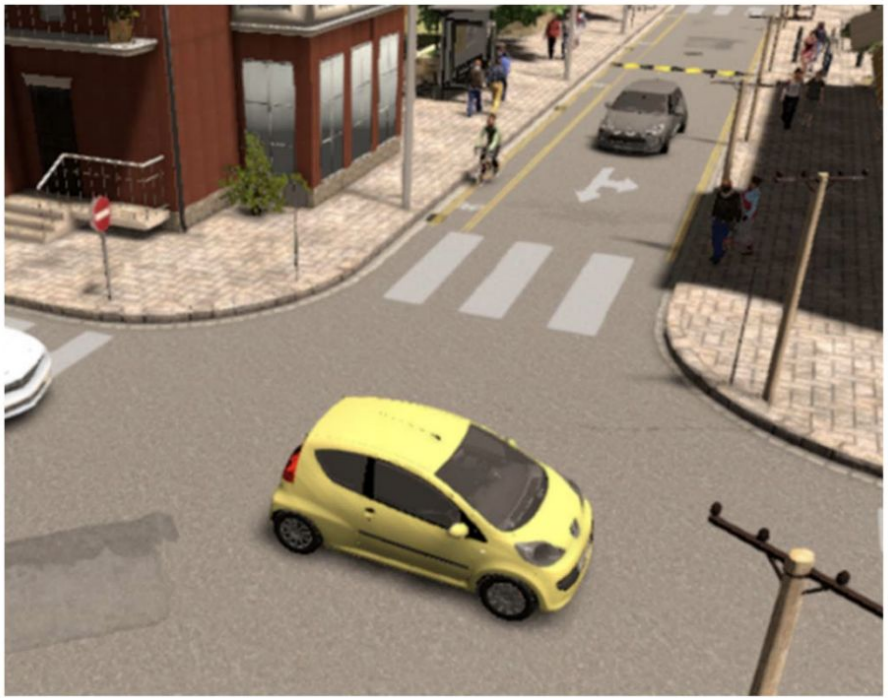
Current trends and future research

- Weakly supervised segmentation
- How to exploit weaker labels to perform semantic segmentation?
  - Image-level annotations https://arxiv.org/abs/1506.03648
  - Point annotations https://arxiv.org/abs/1506.02106
  - Bounding-box annotations https://arxiv.org/abs/1503.01640
  - Scribble annotations https://arxiv.org/abs/1604.05144



Current trends and future research

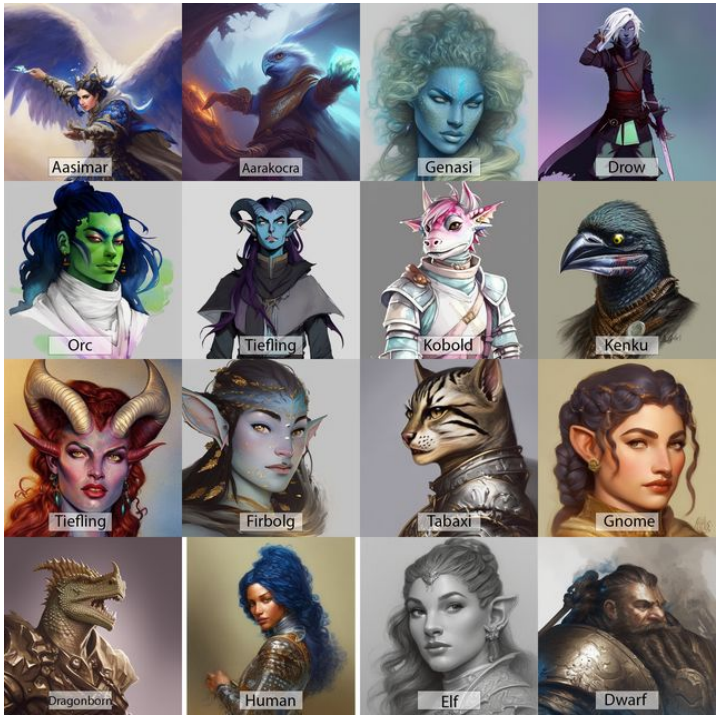
- Sim2Real segmentation
  - Another solution for expensive annotations: **synthetic data**!
  - The total control of the simulation allows for a **potentially infinite** amount of situations, categories, and scenes…
  - **Problem: Domain shift!**



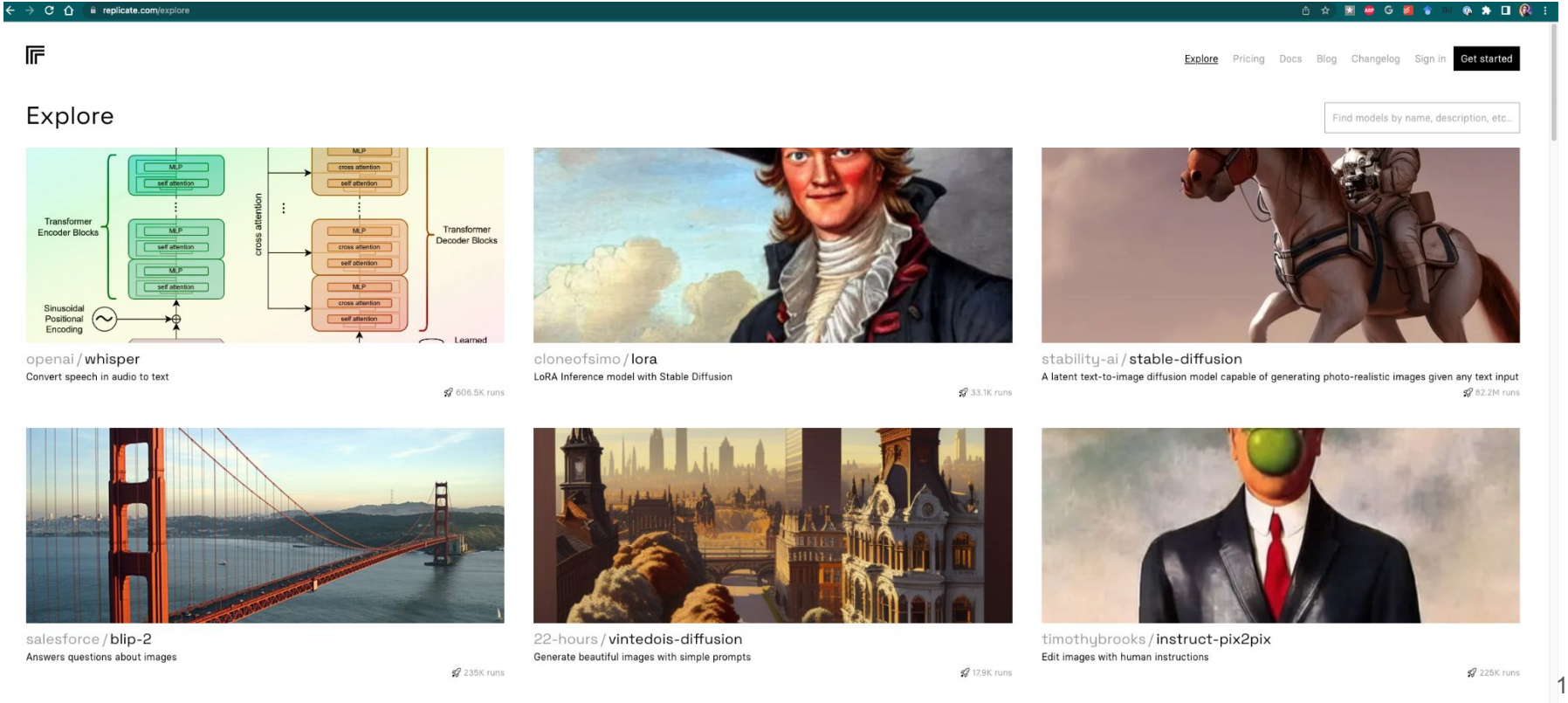
SYNTHIA

Exciting new works

- Use Stable Diffusion or Midjourney to generate images
- Train Semantic Segmentation for free between say different types of characters (**Orc** vs. **Elfs**)
- **Contact me** if interested in these type of projects :)



Exciting new works



Semantic and Instance Segmentation

THANKS FOR YOUR ATTENTION

Issam Laradji

ServiceNow Research

issam.laradji@servicenow.com

Contact me for project and internship opportunities

Acknowledgements:

David Vazquez, Carlos Ventura, German Ros, Pedro Pinheiro and Alberto Garcia-Garcia