Master in
Computer Vision
Barcelona

Semantic and Instance Segmentation

David Vazquez
david.vazquez@servicenow.com

About me

- Research Scientist at **ServiceNow Research,** Montreal, Canada
- Research Lead of the **Low Data Learning** team at ServiceNow Research
- Manager of the **Research Programs** at ServiceNow Research

Formerly:

- Postdoc at UAB and MILA (2013-2016)
- MSc and PhD at UAB (2009-2013)

About me

- Research interests:
  - Low data learning methods, e.g., Few-shot learning, Self-supervision, Active learning, Un/Semi-supervised learning, Continual learning, Domain adaptation, Reinforcement learning, ...
  - Computer vision, e.g., Segmentation, detection, counting, VQA, etc.
  - Graphs, knowledge graphs, and natural language processing

- Master/PhD Thesis supervision and internships
  - Contact with me in case of interest

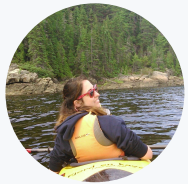
servicenow®

Spanish researchers at Montreal

David Vazquez
ServiceNow

Pau Rodriguez
ServiceNow

Oscar Mañas
Mila/ServiceNow



Adriana Romero
Facebook Research

Michal Drozdal
Facebook Research

Arantxa Casanova
MILA/Facebook

Outline

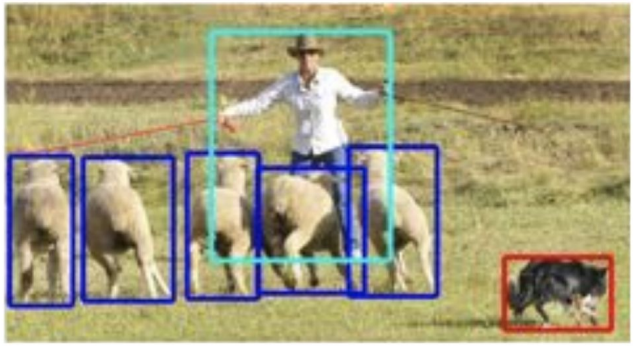
- Introduction to segmentation
- Semantic segmentation
- Instance segmentation
- Panoptic segmentation
- Amodal segmentation
- Referring image segmentation
- Current trends and future research

Outline

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

Introduction to segmentation

(a) image classification



(b) object detection

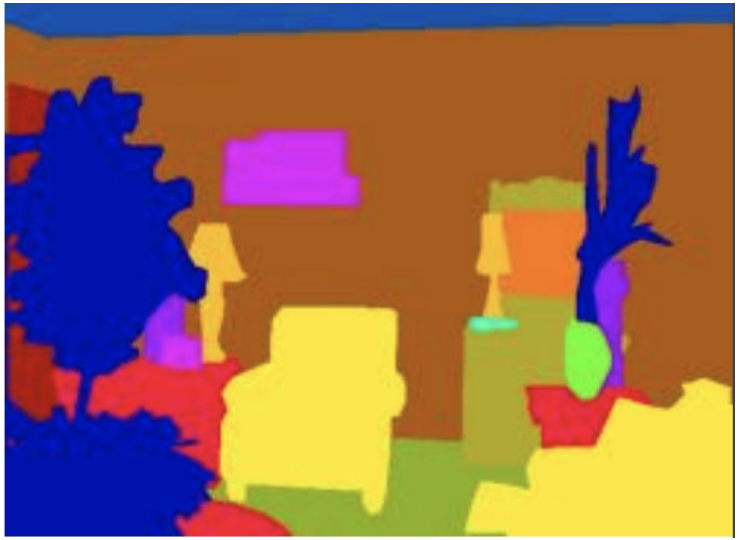
(c) semantic segmentation

(d) instance segmentation

Outline

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

Semantic segmentation: Problem statement



- Give a semantic label to every pixel in an image
- Segmentation and multi-label recognition at once

Semantic segmentation: Problem statement

Find a way to assign a state from the label space $\boldsymbol{L = \{l_1, l_2, \ldots, l_C\}}$, to each one of the elements of a set of random variables $X = \{x_1, x_2, \ldots, x_N\}$.

Each label l_i represents a different class or object (e.g., airplane, car, traffic sign, …).

The set of random variables is usually, but not necessarily, a 2D RGB image (where $N = H \times W$) so the problem becomes a mapping between two tensor spaces:

$$\boldsymbol{f}: \mathbb{R}^{\boldsymbol{H \times W \times 3}} \rightarrow \mathbb{N}^{\boldsymbol{H \times W \times C}}$$

Semantic segmentation: Problem statement

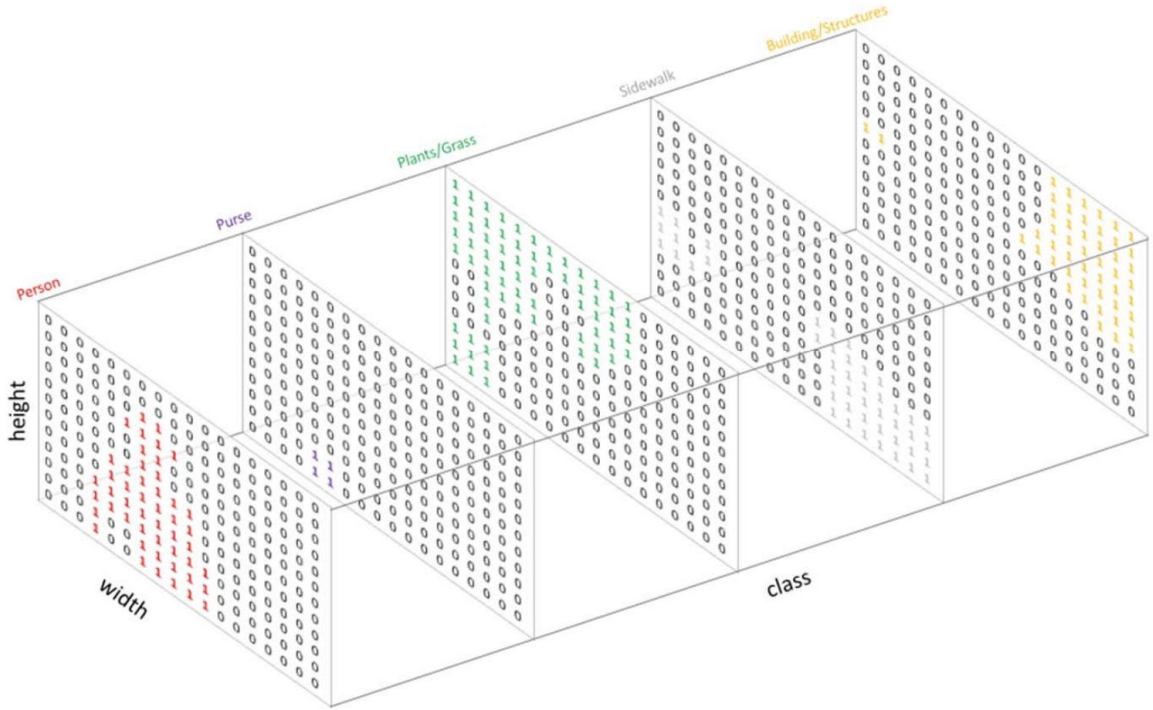


Image source: https://www.jeremyjordan.me/semantic-segmentation/

Semantic segmentation: Problem statement

segmented

1: Person
2: Purse
3: Plants/Grass
4: Sidewalk
5: Building/Structures

3 3 3 3 3 3 3 3 3 3 3 3 5 5 5 5 5 5
3 3 3 3 3 3 3 3 3 3 3 3 5 5 5 5 5 5
3 3 3 3 3 3 1 1 3 3 3 3 5 5 5 5 5 5
3 3 3 3 3 1 1 1 1 3 3 3 5 5 5 5 5 5
3 3 3 3 3 3 1 1 3 3 3 5 5 5 5 5 5 5
5 5 3 3 3 3 1 1 3 3 5 5 5 5 5 5 5 5
4 4 3 4 1 1 1 1 1 1 4 4 4 5 5 5 5 5
4 4 3 4 1 1 1 1 1 1 4 4 4 4 4 5 5 5
4 4 4 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4
3 3 3 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4
3 3 3 1 2 2 1 1 1 1 1 4 4 4 4 4 4 4
3 3 3 1 2 2 1 1 1 1 1 4 4 4 4 4 4 4

Input

Semantic Labels

Image source: https://www.jeremyjordan.me/semantic-segmentation/

Semantic segmentation: applications

- Autonomous driving

Image source: https://medium.com/intro-to-artificial-intelligence/semantic-segmentation-udaitys-self-driving-car-engineer-nanodegree-c01eb6eaf9d

Semantic segmentation: applications

- Medical imaging

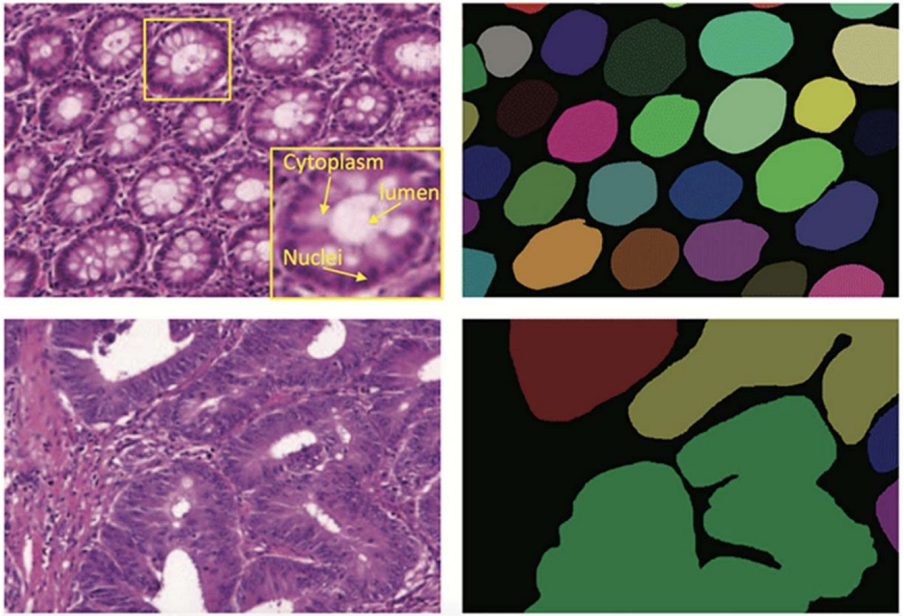


Image source: NVIDIA Blog: GPU-Powered Deep Learning Being Used to Speed Colon Cancer Diagnosis

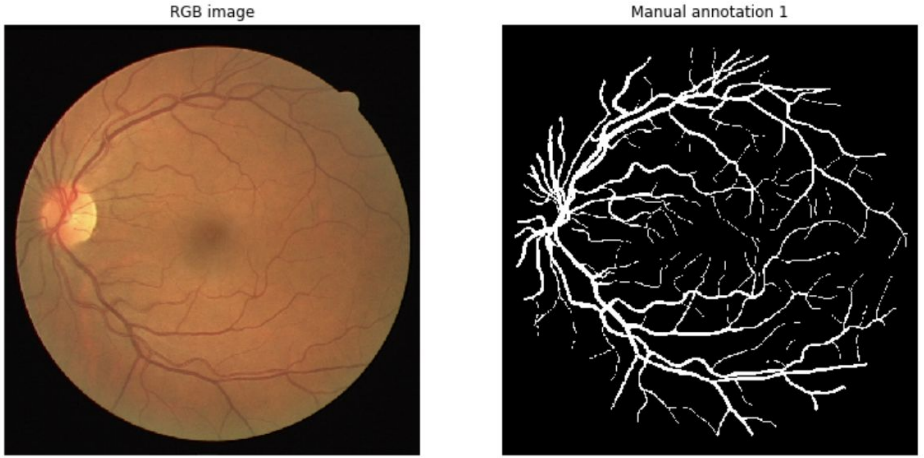
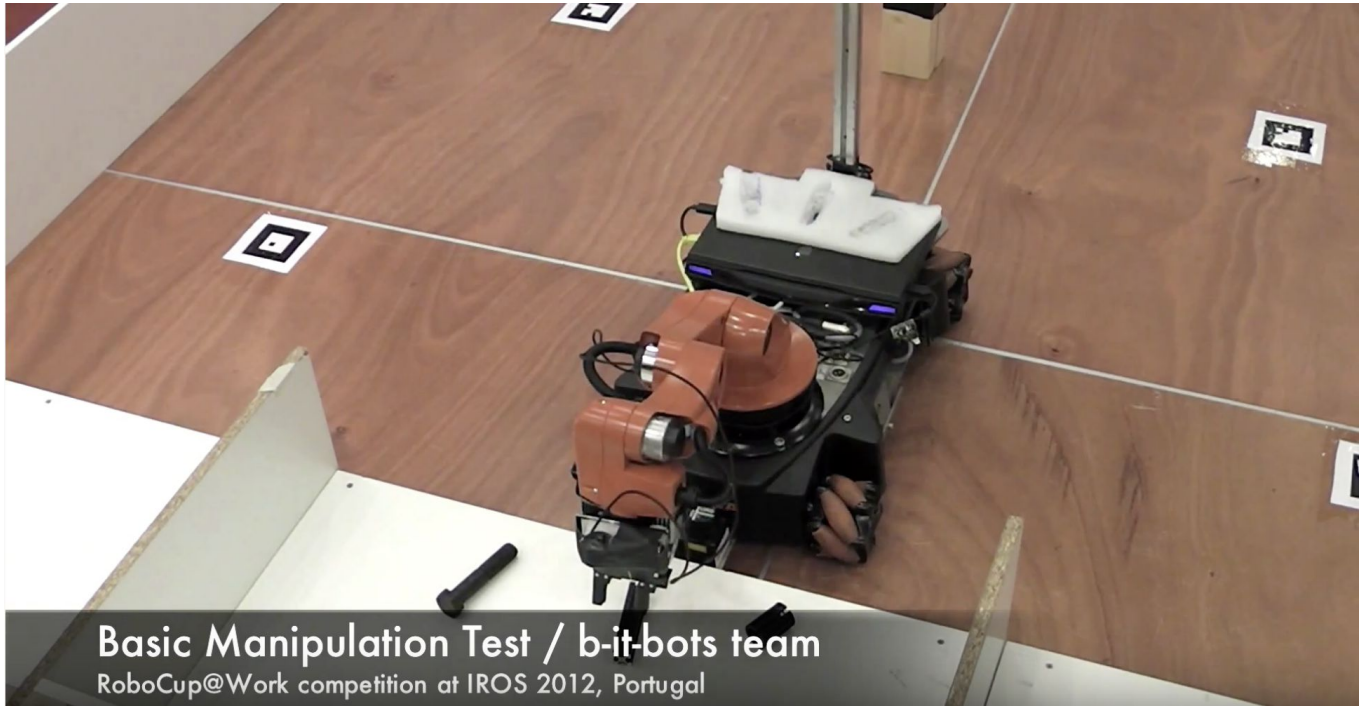


Image source: DRIVE Digital Retinal Image Vessel Extraction

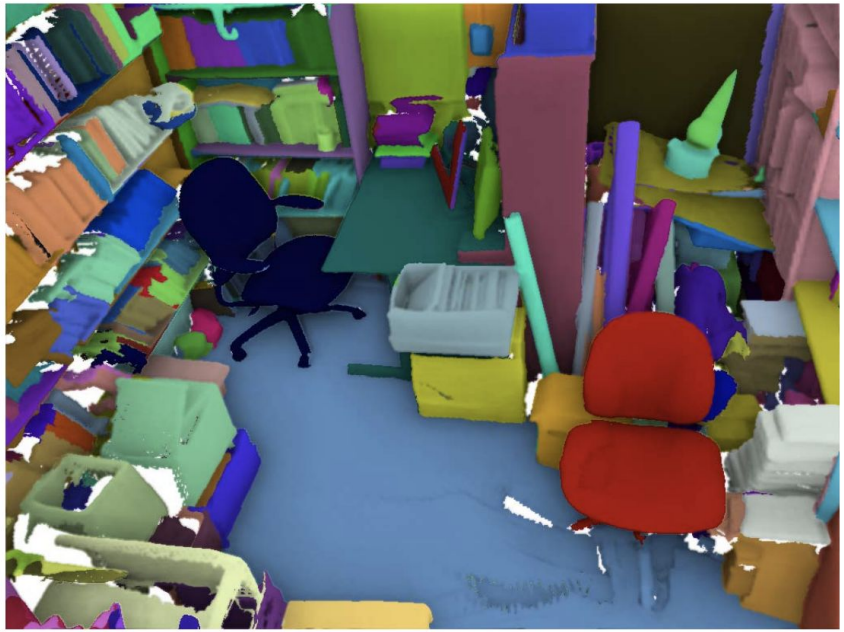
Semantic segmentation: applications

- Robotic applications



Semantic segmentation: applications

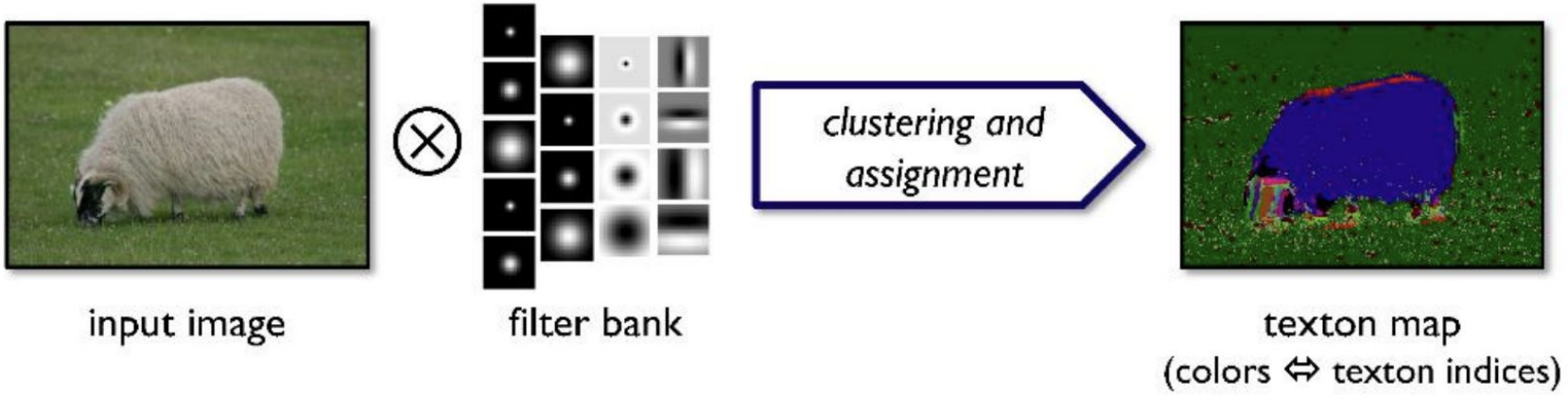
- Scene understanding



Images sourced from: SceneNN: A Scene Meshes Dataset with aNNotations

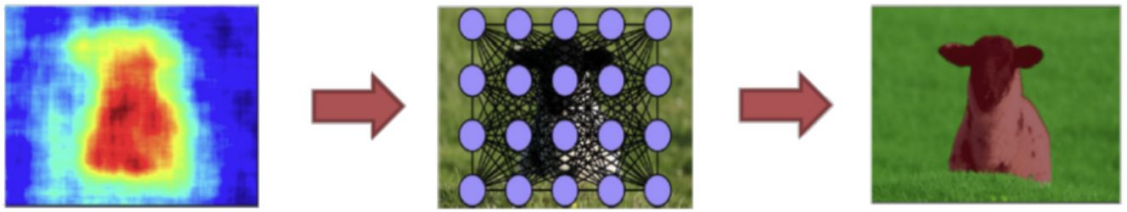
Semantic segmentation: traditional approaches

- Semantic segmentation before deep learning
  - Relying on conditional random field (CRF)
  - Operating on pixels or superpixels
  - Incorporating local evidence in unary potentials
  - Interactions between label assignments



Semantic segmentation: traditional approaches

- Conditional Random Field?
  - CRF is a probabilistic framework for labeling and segmenting structured data
  - CRF is still applied nowadays as a post-processing technique
  - Basic ideas:
    - Nearby pixels more likely to have same label
    - Pixels with similar color/texture/… more likely to have same label
    - Pixels surrounded by “river” label more likely to be a boat than a car
    - Refine results by iterations



Coarse output from the pixel-wise classifier

MRF/CRF modelling

Output after the CRF inference

Semantic segmentation: traditional approaches

- Superpixels?
  - A superpixel can be defined as a group of pixels that share common characteristics (like pixel intensity)
  - Example technique: SLIC (Simple Linear Iterative Clustering)
    - Clustering based on the color similarity and spatial proximity

Semantic segmentation: traditional approaches

Image source: PASCAL VOC 2012

Pre-segmentation
+
Hand-crafted Descriptors

Local Classification
+
Global Context

Semantic segmentation: traditional approaches

Pre-segmentation
+
Hand-crafted Descriptors

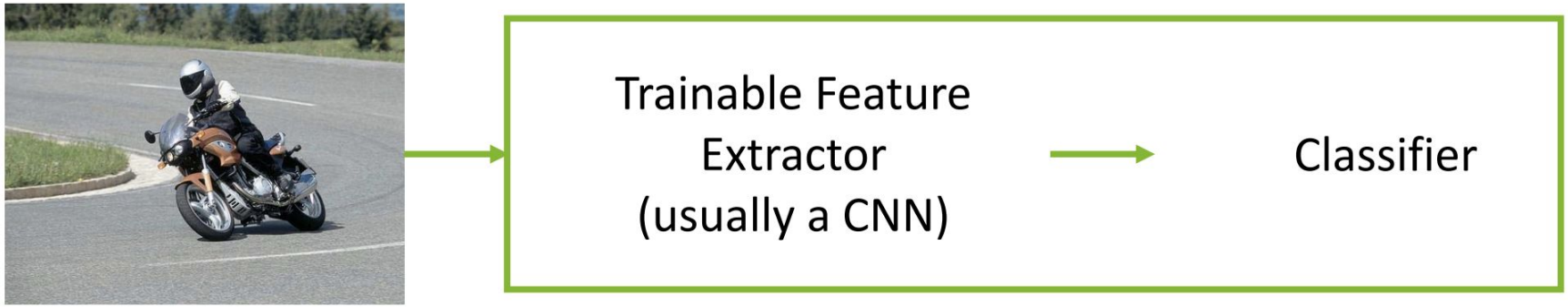
Local Classification
+
Global Context

Image source: PASCAL VOC 2012

- Drawbacks:
  - Pre-segmentation methods are not perfect and are hard to tune
  - Hand-crafter descriptors
  - Global context integration (e.g. CRF) is computationally expensive

Semantic segmentation: traditional approaches

- Intermediate solution:



  - Pre-segmentation and description are replaced by a trainable feature extractor
  - The classifier is trained jointly with the previous stage
  - **… but how do we move from classification to segmentation?**

Semantic segmentation: traditional approaches

- Intermediate solution: patch approach

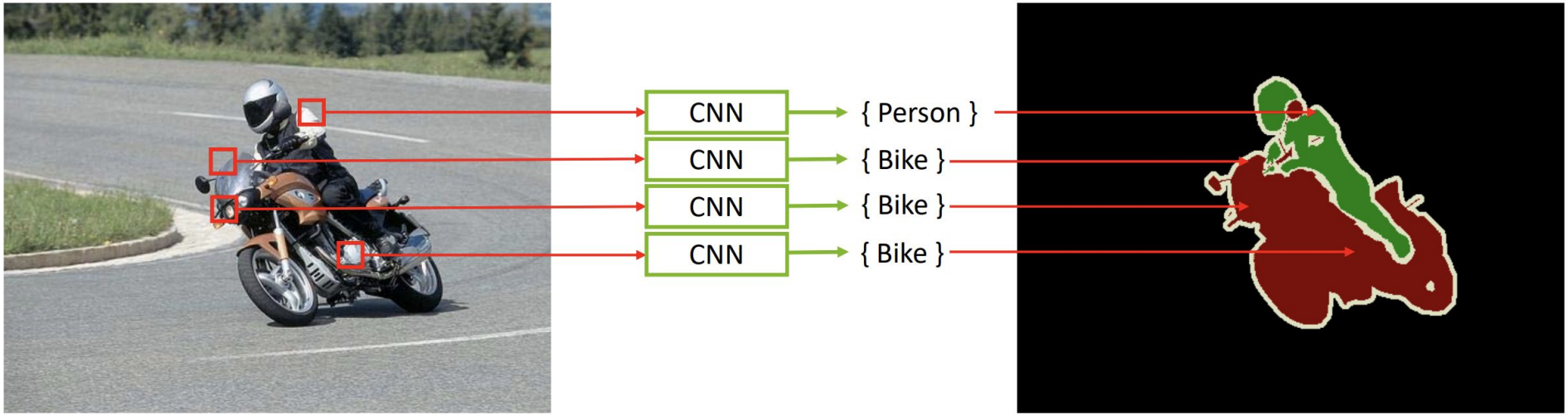


Image source: PASCAL VOC 2012

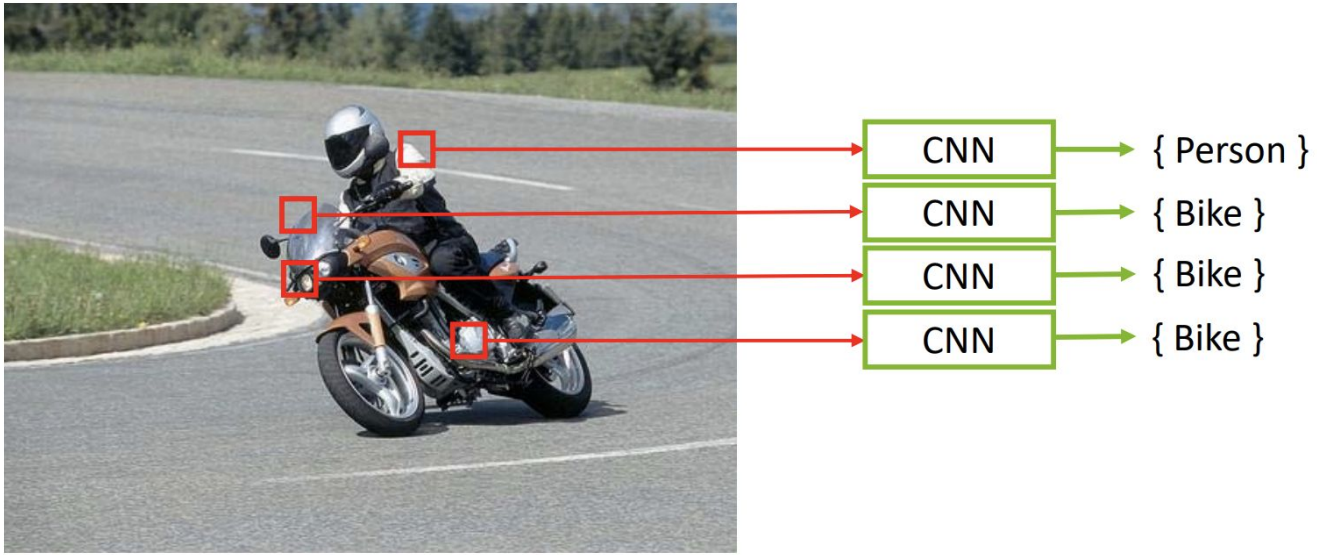
Semantic segmentation: traditional approaches

- Intermediate solution: patch approach

Basic and intuitive

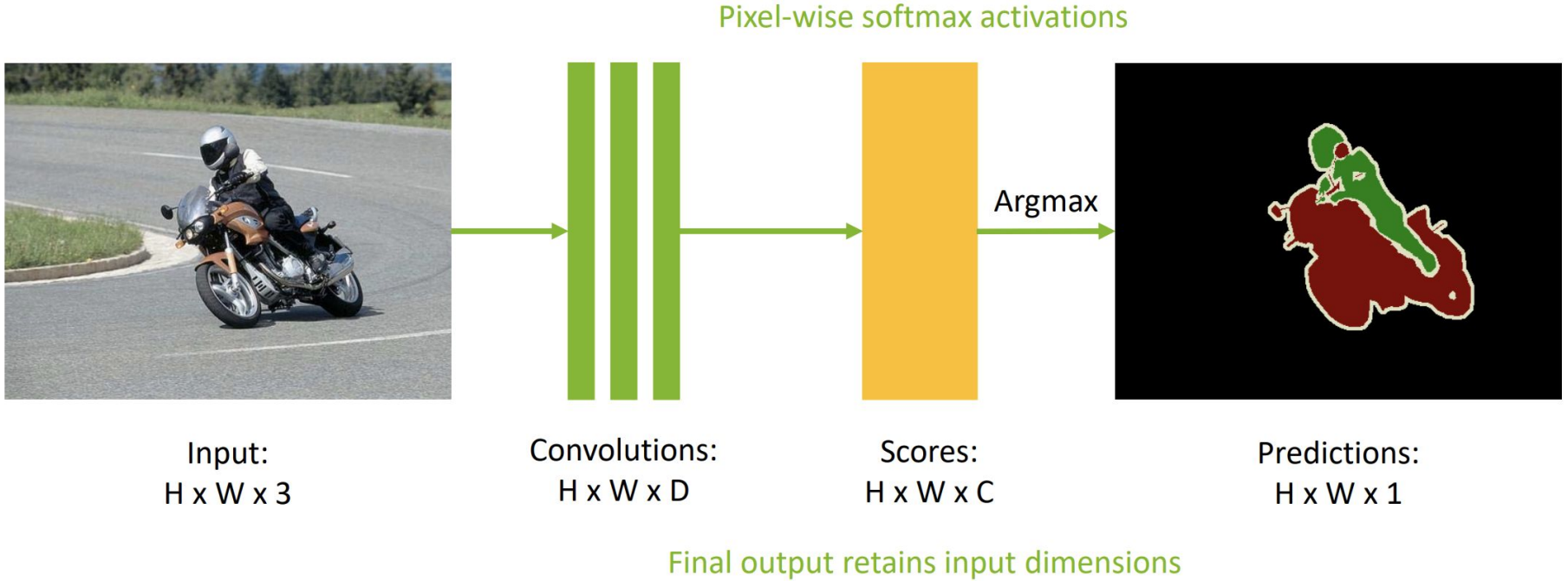
Cheap (memory)

Inefficient training does not reuse shared features between overlapping patches!



Semantic segmentation: traditional approaches

- Intermediate solution: full-image approach



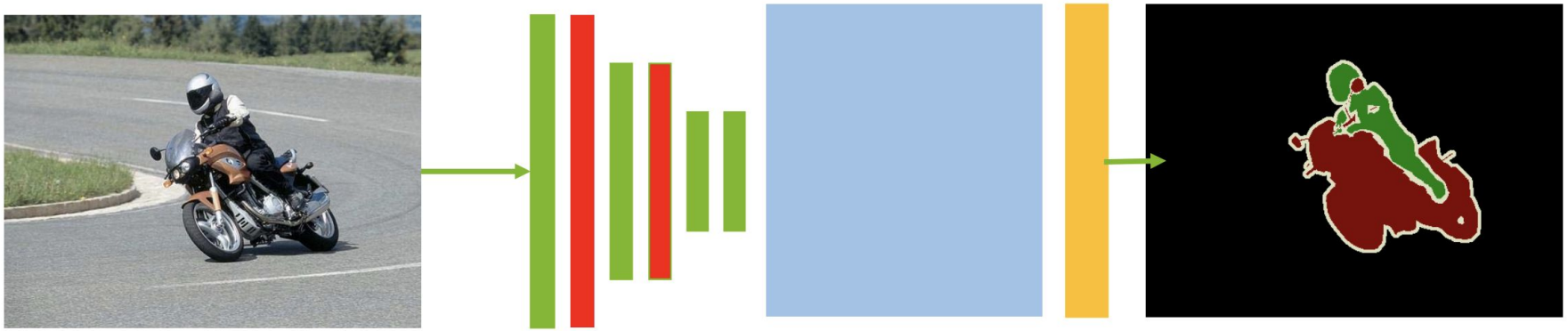
Semantic segmentation: traditional approaches

- Intermediate solution: full-image approach

Convolutions at original image resolution are expensive and impractical in most cases!

Semantic segmentation: traditional approaches

- Intermediate solution: full-image approach



Downsampling with pooling and convolutions renders the network more efficient!

But how do we implement upsampling?

Semantic segmentation: traditional approaches

- Intermediate solution: full-image approach
  - Bilinear interpolation

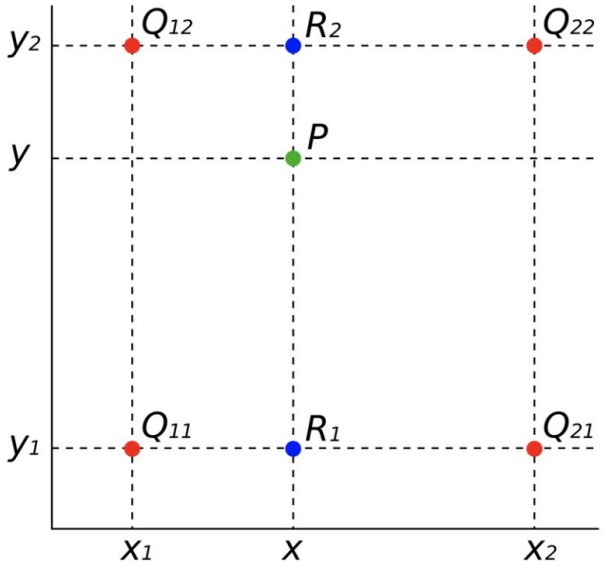


Image source: Wikipedia

Semantic segmentation: traditional approaches

- Intermediate solution: full-image approach
  - Unpooling

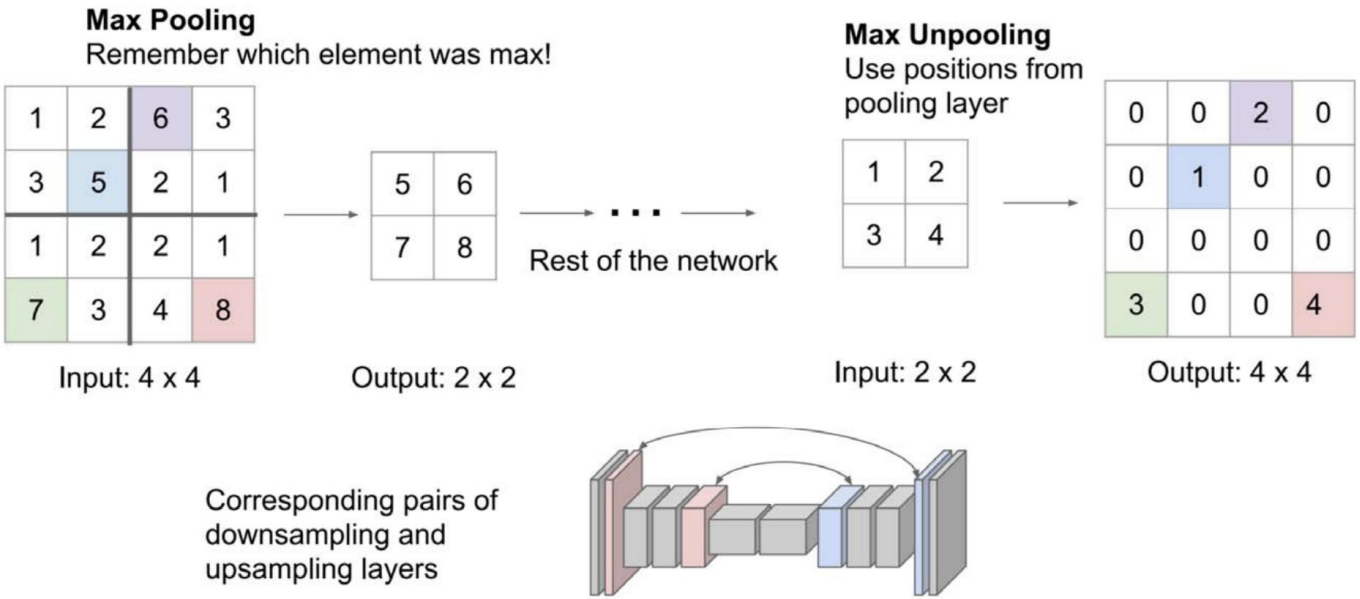
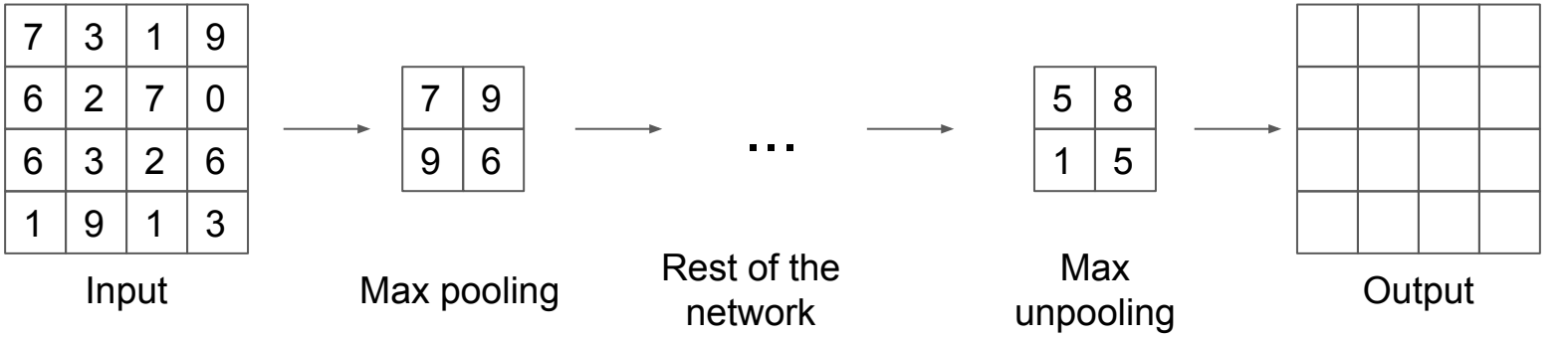
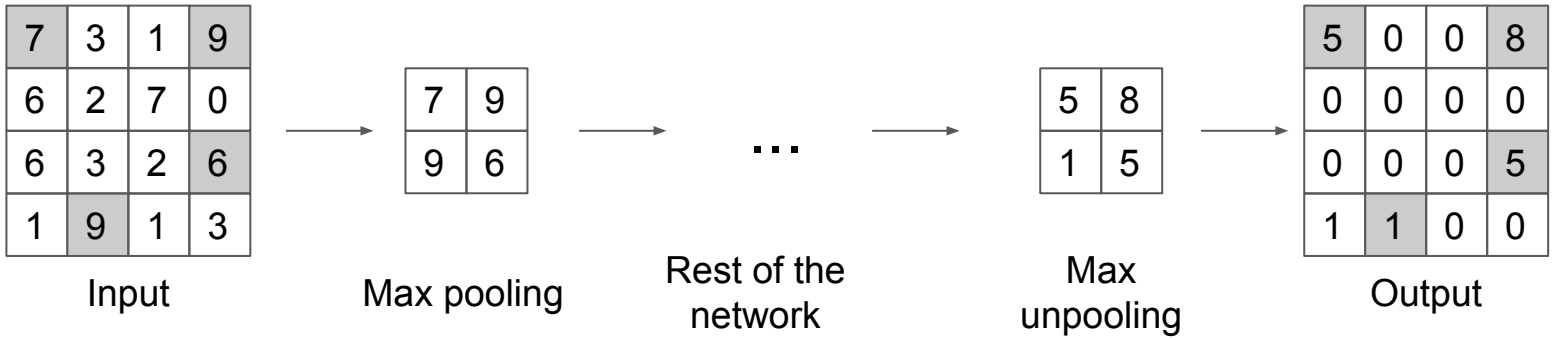


Image source: Stanford CS231 Lecture 11


7 3 1 9
6 2 7 0
6 3 2 6
1 9 1 3
Input
7 9
9 6
Max pooling
...
Rest of the network
5 8
1 5
Max unpooling
Output


7 3 1 9
6 2 7 0
6 3 2 6
1 9 1 3
Input
7 9
9 6
Max pooling
...
Rest of the network
5 8
1 5
Max unpooling
5 0 0 8
0 0 0 0
0 0 0 5
1 1 0 0
Output

Semantic segmentation: traditional approaches

- Intermediate solution: full-image approach
  - Transposed convolution

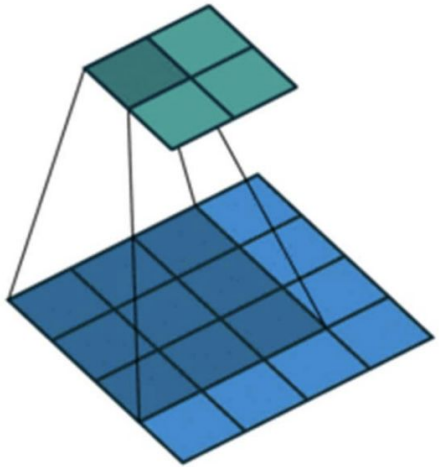


Image source: DeepLearning.net: Convolution Arithmetic

Convolution of

a 3x3 kernel
on a 4x4 input
with unitary stride
and no padding

Produces a 2x2 output!

Semantic segmentation: traditional approaches

- Intermediate solution: full-image approach
  - Transposed convolution

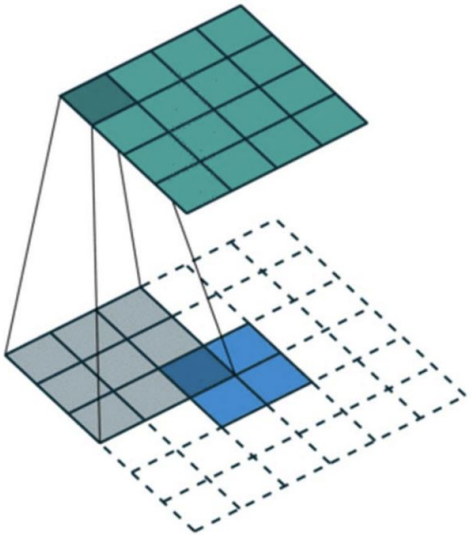


Image source: DeepLearning.net: Convolution Arithmetic

Transposed convolution of

a 3x3 kernel
on a 2x2 input
with unitary stride
and 2x2 pading

Produces a 4x4 output!

Usually initialized as a bilinear transform!

Semantic segmentation: modern approaches

- Deep learning based approaches

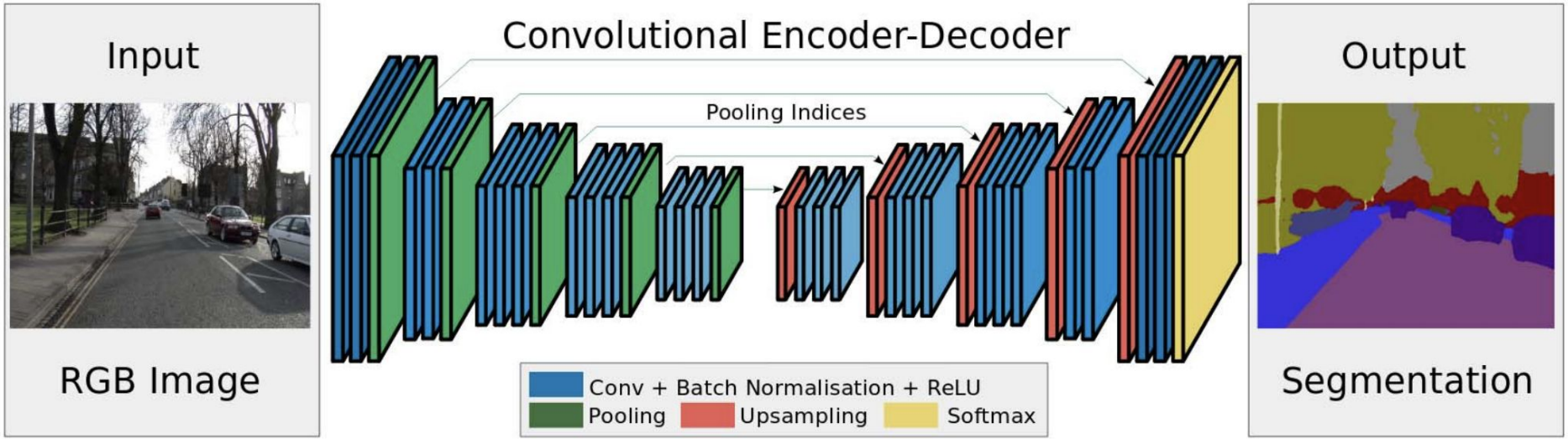
SEGNET · FCN · U-NET

DEEPLAB · PSPNET · FPN

REFINENET · PARSENET

Semantic segmentation: modern approaches

- SegNet [arxiv2015, PAMI2017] (>4K citations)



SegNet encoder/decoder followed by softmax for pixel-wise classification

Image sourced from Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation."

Semantic segmentation: modern approaches

- SegNet [arxiv2015, PAMI2017]
  - PAMI 2017: https://ieeexplore.ieee.org/abstract/document/7803544
  - arxiv 2015: https://arxiv.org/abs/1505.07293

Semantic segmentation: modern approaches

- Fully Convolutional Network (FCN) [CVPR2015, PAMI2016] (>14K citations)

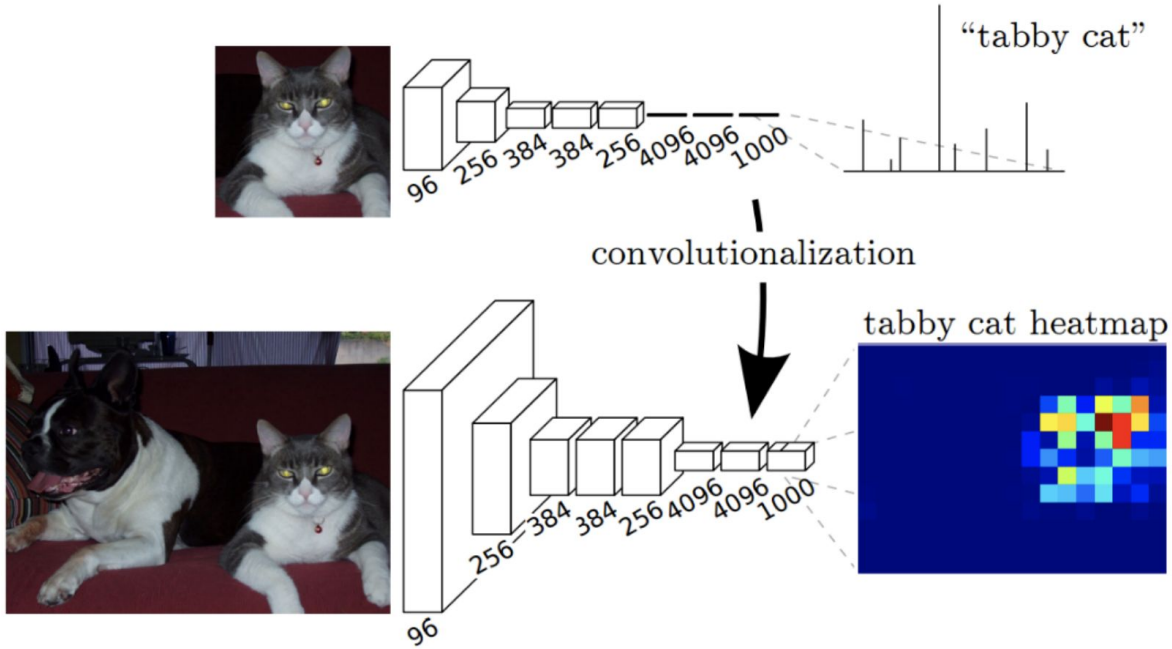


Image sourced from Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation."

Semantic segmentation: modern approaches

- Fully Convolutional Network (FCN) [CVPR2015, PAMI2016] (>14K citations)

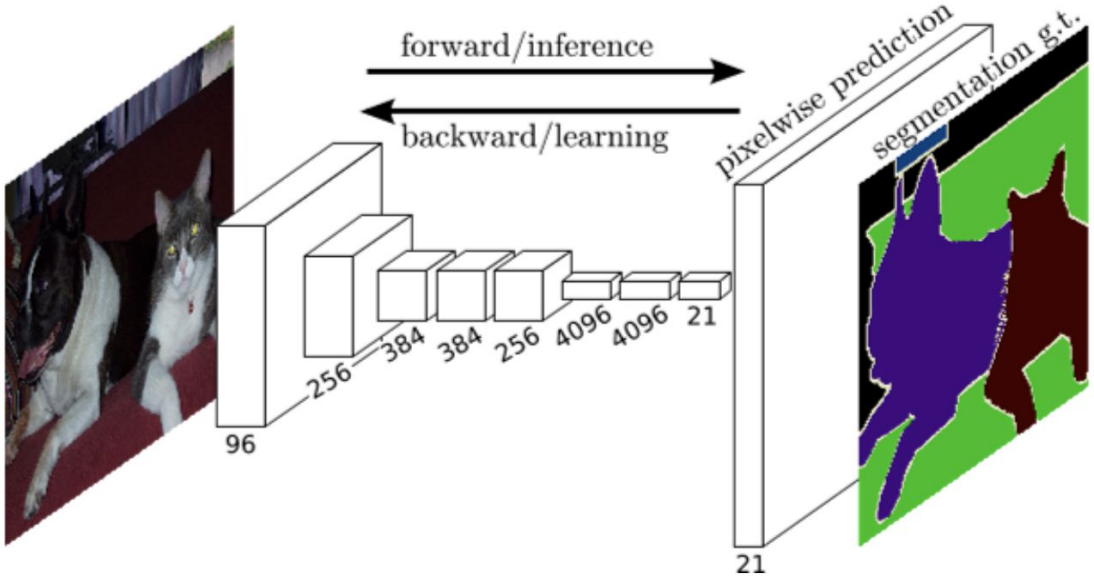
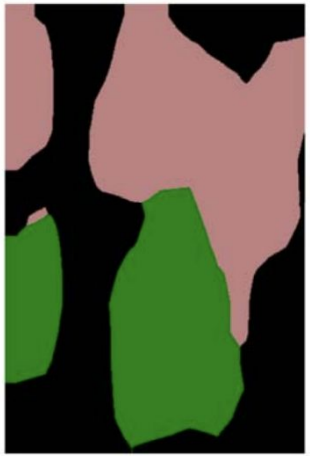


Image sourced from Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation."

Semantic segmentation: modern approaches

- Fully Convolutional Network (FCN) [CVPR2015, PAMI2016] (>14K citations)

FCN-32s



Encoder reduces input resolution by a factor or 32x!!!

Decoder struggles to produce good segmentations!

How can we improve that?

Ground truth

Image sourced from Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation."

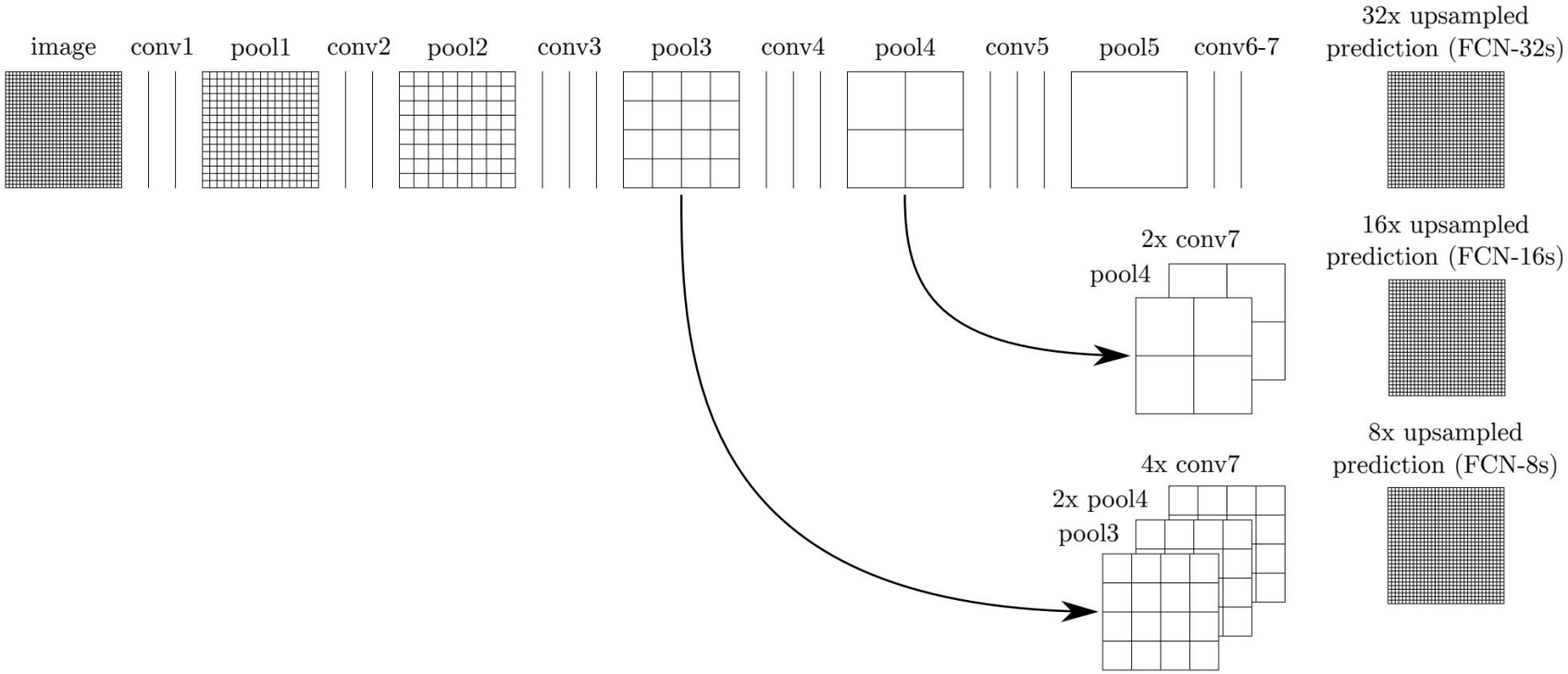
Semantic segmentation: modern approaches

- Fully Convolutional Network (FCN) [CVPR2015, PAMI2016] (>14K citations)

"Semantic segmentation faces an inherent tension between semantics and location: global information resolves **what** while local information resolves **where**... Combining fine layers and coarse layers lets the model make local predictions that respect global structure."

Semantic segmentation: modern approaches

- Fully Convolutional Network (FCN) [CVPR2015, PAMI2016] (>14K citations)



Semantic segmentation: modern approaches

- Fully Convolutional Network (FCN) [CVPR2015, PAMI2016] (>14K citations)

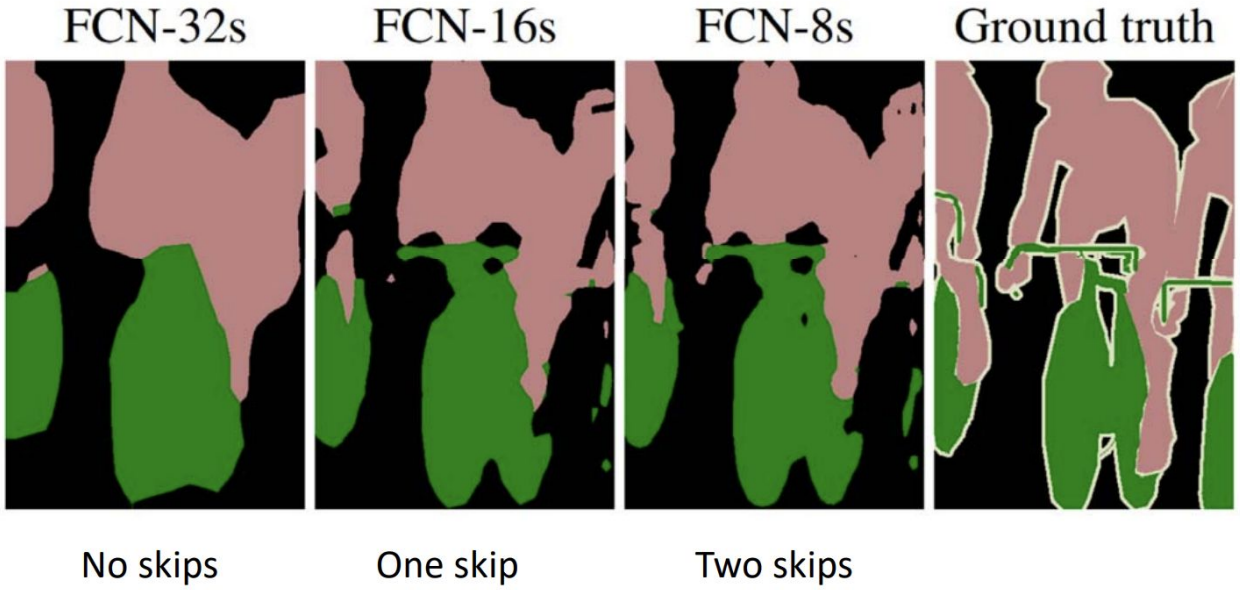


Image sourced from Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation."

Semantic segmentation: modern approaches

- Fully Convolutional Network (FCN) [CVPR2015, PAMI2016] (>14K citations)
  - CVPR 2015: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Long_Fully_Convolutional_Networks_2015_CVPR_paper.pdf
  - PAMI 2016: https://ieeexplore.ieee.org/document/7478072

Semantic segmentation: modern approaches

- U-Net [MICCAI2015, PAMI2016] (>11K citations)

  Improve upon FCN

  Expand decoder capacity

  Contracting path (context)

  Expanding path (precise)

  Heavy data augmentation

  Widely used for medical images

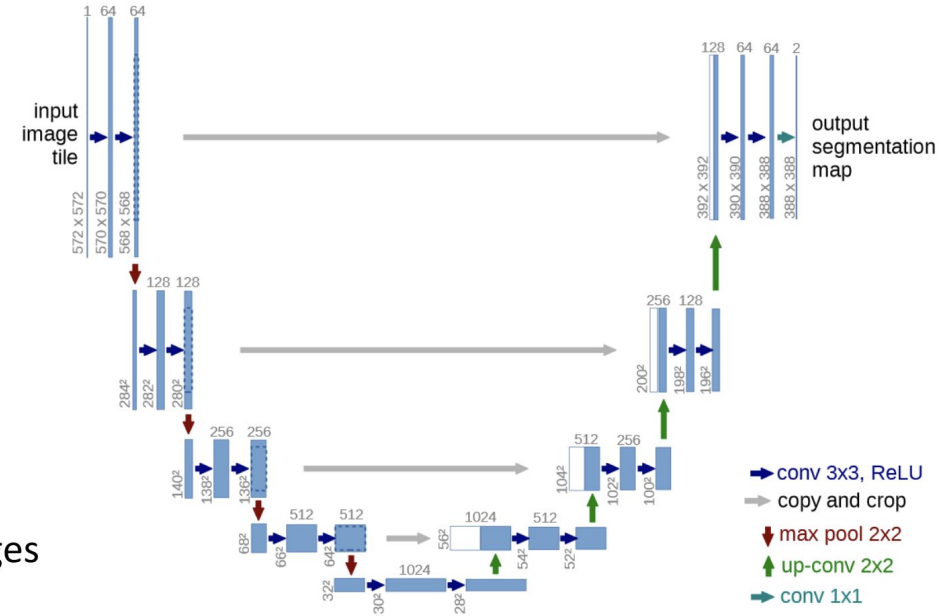


Image sourced from Ronneberger, O., Fischer, P., & Brox, T. “U-net: Convolutional networks for biomedical image segmentation”

Semantic segmentation: modern approaches

- U-Net [MICCAI2015, PAMI2016] (>11K citations)

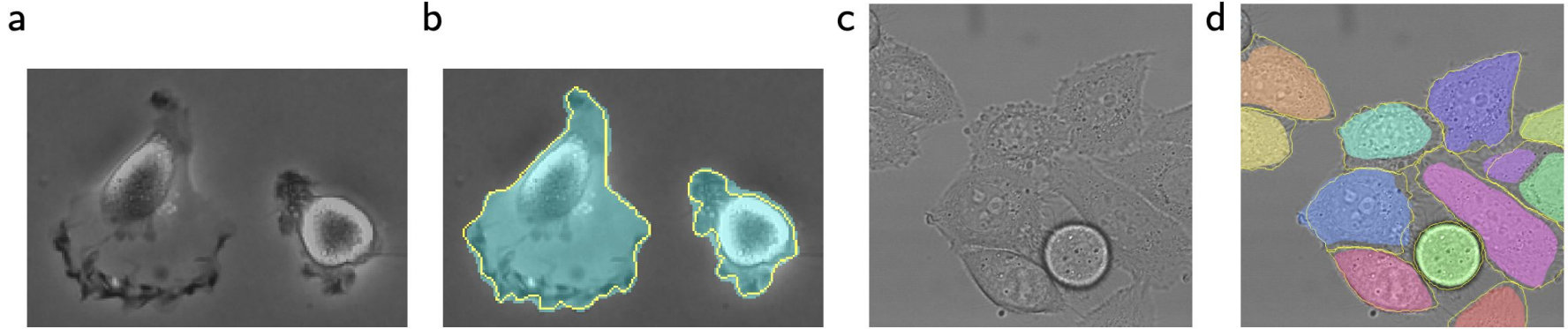


Fig. 4. Result on the ISBI cell tracking challenge. (**a**) part of an input image of the "PhC-U373" data set. (**b**) Segmentation result (cyan mask) with manual ground truth (yellow border) (**c**) input image of the "DIC-HeLa" data set. (**d**) Segmentation result (random colored masks) with manual ground truth (yellow border).

Semantic segmentation: modern approaches

- U-Net [MICCAI2015, PAMI2016] (>11K citations)
  - MICCAI 2015: https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28
  - arxiv version: https://arxiv.org/abs/1505.04597

Semantic segmentation: modern approaches

- DeepLab [ICLR2015, PAMI2017] (>4K citations)

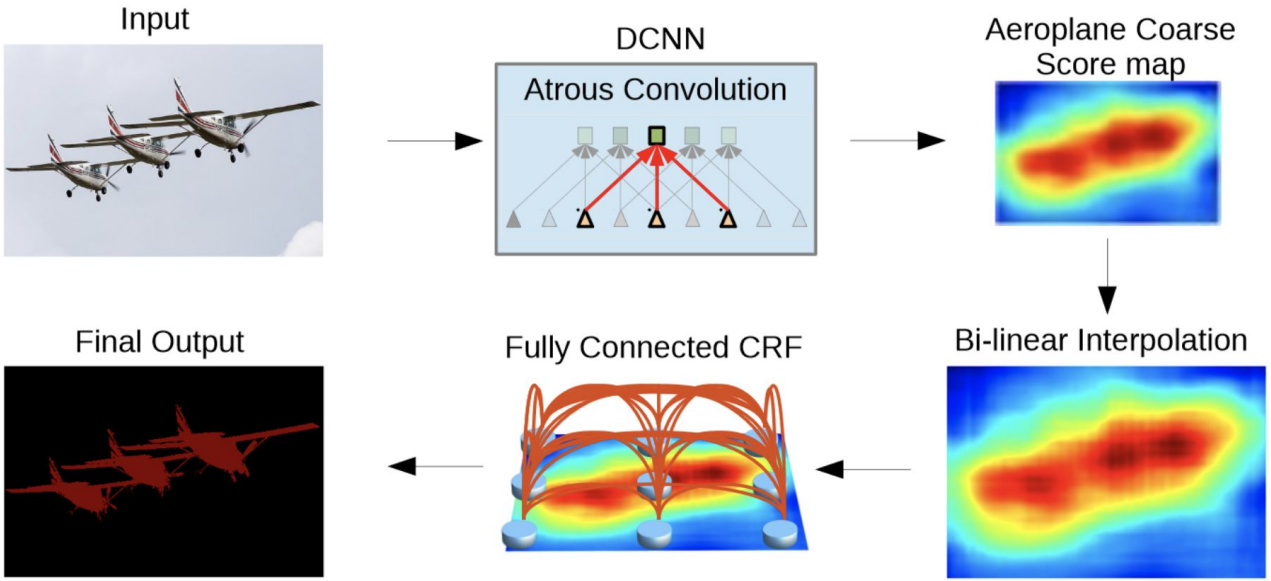


Image sourced from Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L.. “DeepLab: Semantic Image Segmentation withDeep Convolutional Nets, Atrous Convolution,and Fully Connected CRFs”

Semantic segmentation: modern approaches

- DeepLab [ICLR2015, PAMI2017] (>4K citations)

  Pooling loses spatial information

  But if we don’t pool the receptive field becomes too small

  That leads to bad performance

  Use dilated convolutions instead!

  Widen the receptive field

  While avoiding spatial resolution coarsening!

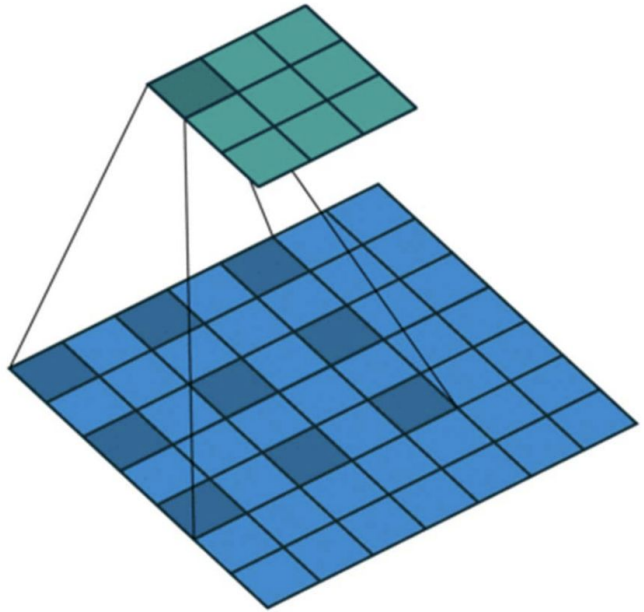
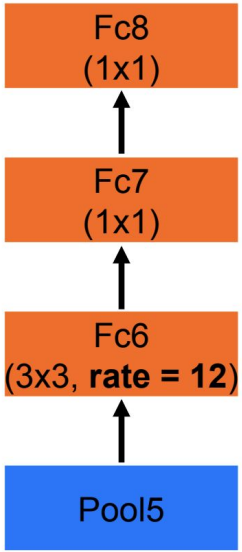


Image source: DeepLearning.net: Convolution Arithmetic

Semantic segmentation: modern approaches

- DeepLab [ICLR2015, PAMI2017] (>4K citations)



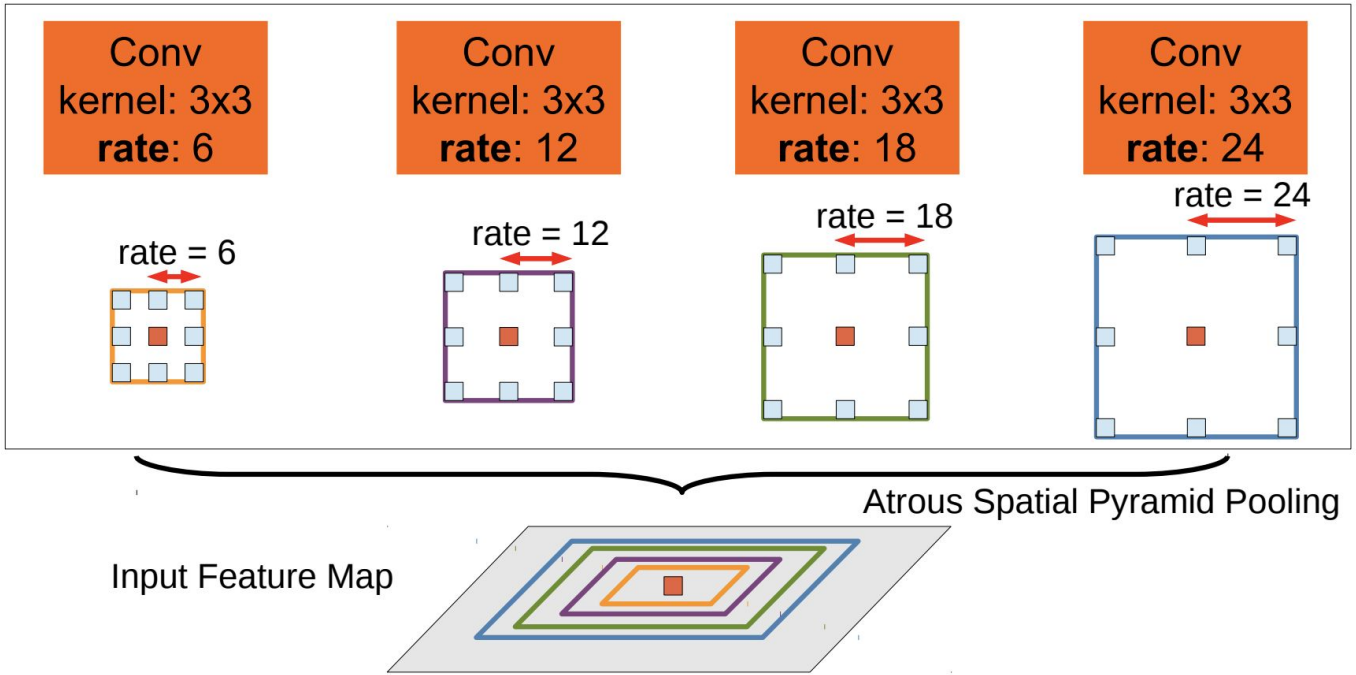
DeepLab-LargeFOV

Kernel	Rate	FOV	Params	Speed	bef/aft CRF
7×7	4	224	134.3M	1.44	64.38 / 67.64
4×4	4	128	65.1M	2.90	59.80 / 63.74
4×4	8	224	65.1M	2.90	63.41 / 67.14
3×3	12	224	20.5M	4.84	62.25 / 67.64

TABLE 1: Effect of Field-Of-View by adjusting the kernel size and atrous sampling rate r at 'fc6' layer. We show number of model parameters, training speed (img/sec), and *val* set mean IOU before and after CRF. DeepLab-LargeFOV (kernel size 3×3, $r = 12$) strikes the best balance.

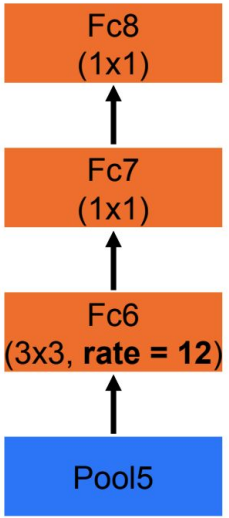
Semantic segmentation: modern approaches

- DeepLab [ICLR2015, PAMI2017] (>4K citations)

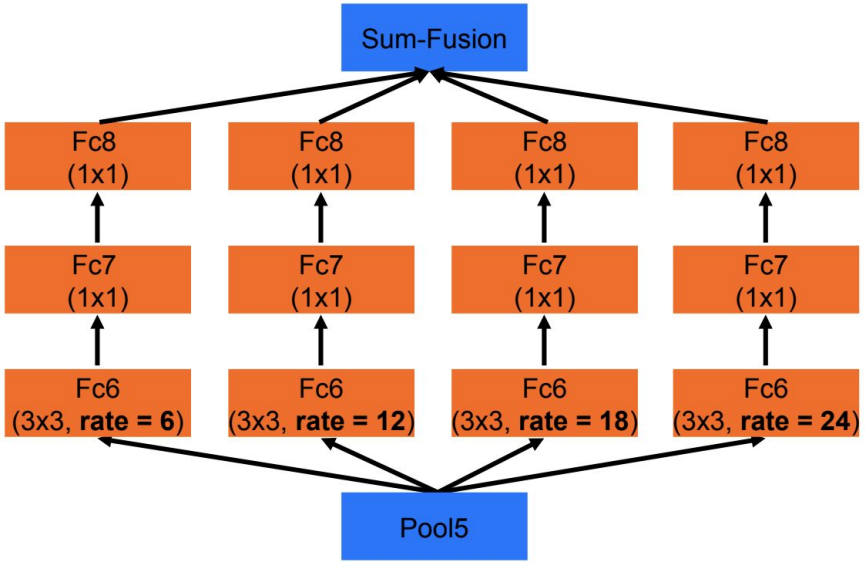


Semantic segmentation: modern approaches

- DeepLab [ICLR2015, PAMI2017] (>4K citations)



(a) DeepLab-LargeFOV



(b) DeepLab-ASPP

Semantic segmentation: modern approaches

- DeepLab [ICLR2015, PAMI2017] (>4K citations)

MSC	COCO	Aug	LargeFOV	ASPP	CRF	mIOU
						68.72
✓						71.27
✓	✓					73.28
✓	✓	✓				74.87
✓	✓	✓	✓			75.54
✓	✓	✓		✓		76.35
✓	✓	✓		✓	✓	77.69

TABLE 4: Employing ResNet-101 for DeepLab on PASCAL VOC 2012 *val* set. **MSC**: Employing mutli-scale inputs with max fusion. **COCO**: Models pretrained on MS-COCO. **Aug**: Data augmentation by randomly rescaling inputs.

Semantic segmentation: modern approaches

- DeepLab [ICLR2015, PAMI2017] (>4K citations)

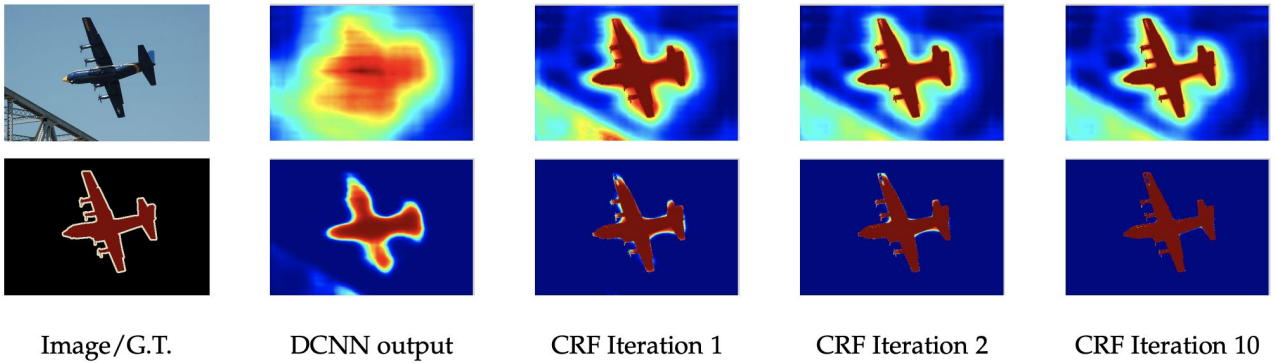


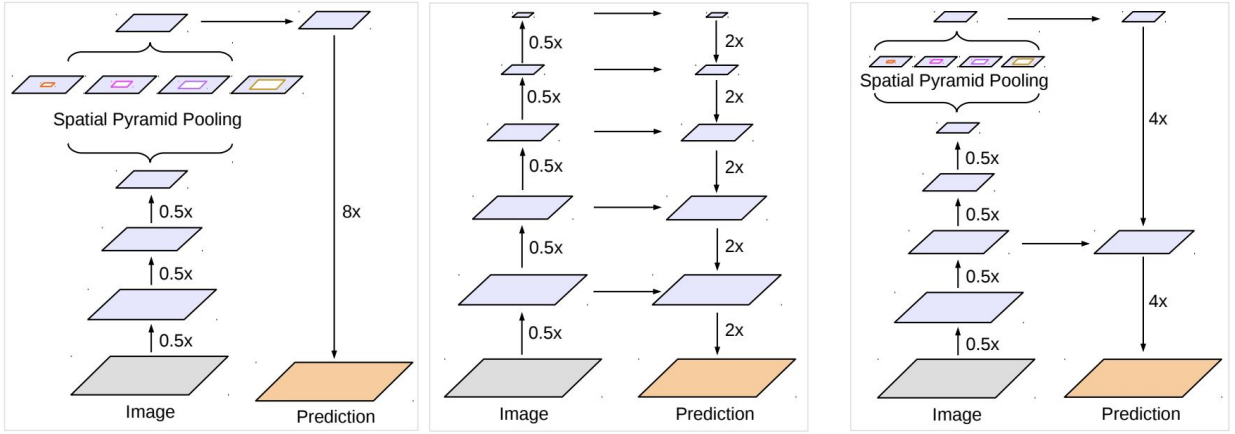
Fig. 5: Score map (input before softmax function) and belief map (output of softmax function) for Aeroplane. We show the score (1st row) and belief (2nd row) maps after each mean field iteration. The output of last DCNN layer is used as input to the mean field inference.

Semantic segmentation: modern approaches

- DeepLab [ICLR2015, PAMI2017] (>4K citations)
  - PAMI 2017: https://ieeexplore.ieee.org/abstract/document/7913730
  - arxiv version: https://arxiv.org/pdf/1606.00915.pdf

Semantic segmentation: modern approaches

- DeepLab v3+ [ECCV2018] (>1K citations)

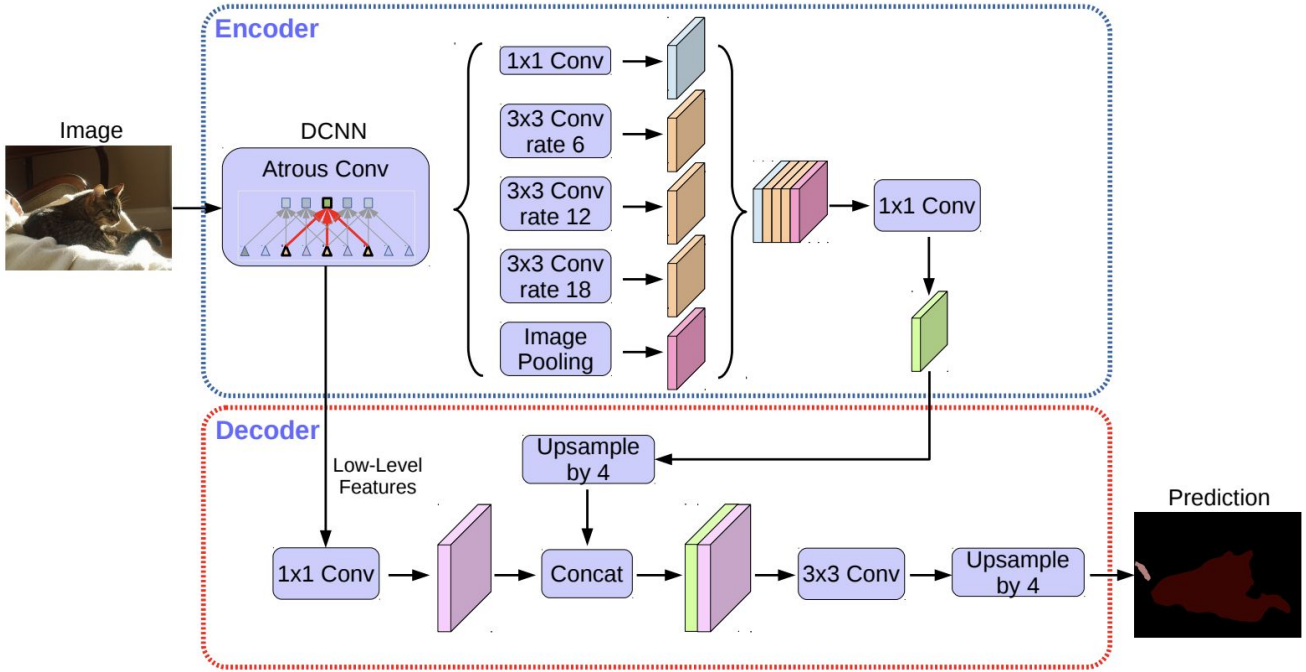


(a) Spatial Pyramid Pooling (b) Encoder-Decoder (c) Encoder-Decoder with Atrous Conv

Fig. 1. We improve DeepLabv3, which employs the spatial pyramid pooling module (a), with the encoder-decoder structure (b). The proposed model, DeepLabv3+, contains rich semantic information from the encoder module, while the detailed object boundaries are recovered by the simple yet effective decoder module. The encoder module allows us to extract features at an arbitrary resolution by applying atrous convolution.

Semantic segmentation: modern approaches

- DeepLab v3+ [ECCV2018] (>1K citations)

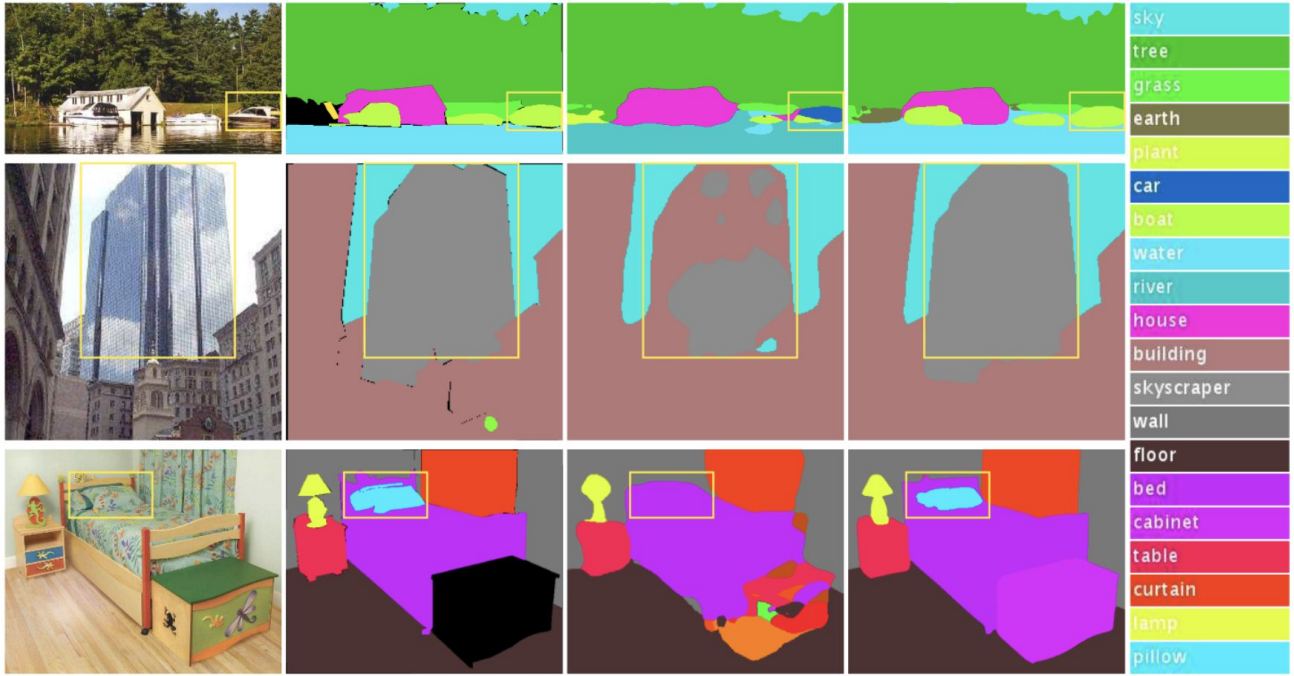


Semantic segmentation: modern approaches

- DeepLab v3+ [ECCV2018] (>1K citations)
  - ECCV 2018: http://openaccess.thecvf.com/content_ECCV_2018/papers/Liang-Chieh_Chen_Encoder-Decoder_with_Atrous_ECCV_2018_paper.pdf

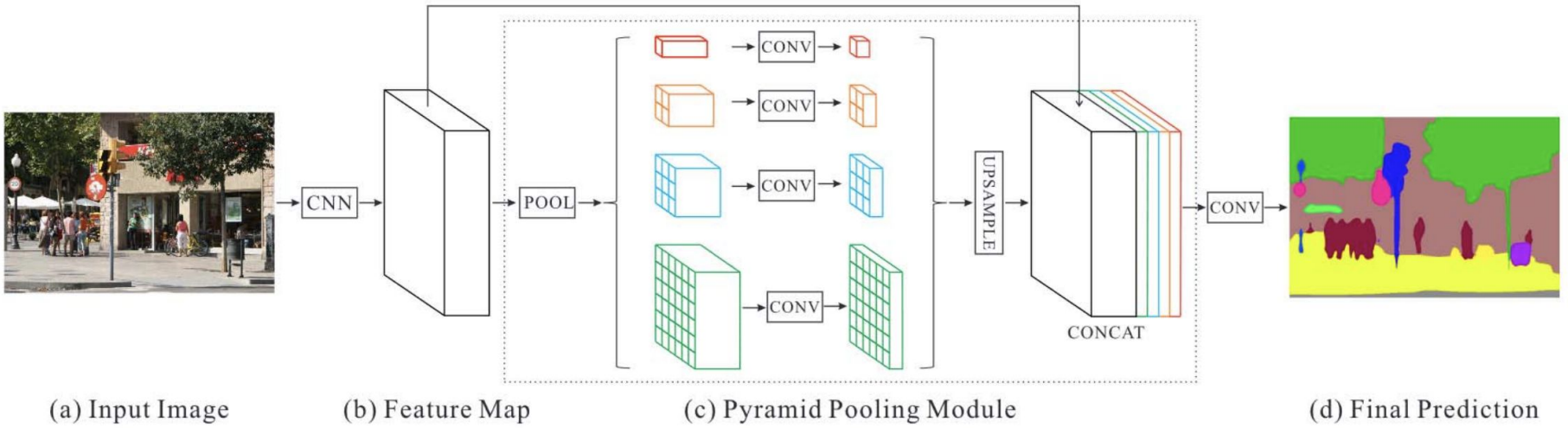
Semantic segmentation: modern approaches

- Pyramid Scene Parsing Network (PSPNet) [CVPR2017] (>2K citations)



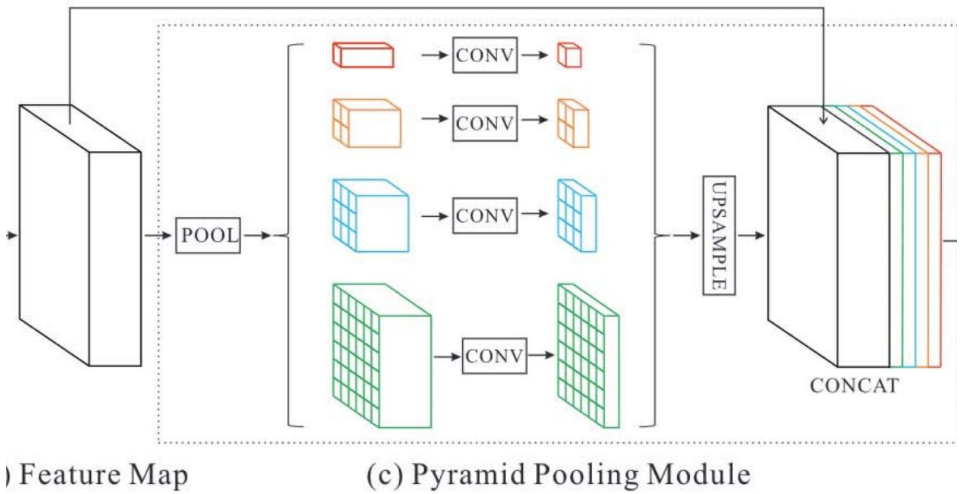
Semantic segmentation: modern approaches

- Pyramid Scene Parsing Network (PSPNet) [CVPR2017] (>2K citations)



Semantic segmentation: modern approaches

- Pyramid Scene Parsing Network (PSPNet) [CVPR2017] (>2K citations)
  - Patterns are sometimes relationships between small and large objects
  - Hard to account for both with classical architectures
  - Pyramid pooling module provides global-scene priors over 4 different scales:
    - over whole input
    - over 2x2 areas of input
    - over 3x3 areas of input
    - over 6x6 areas of input



Semantic segmentation: modern approaches

- Pyramid Scene Parsing Network (PSPNet) [CVPR2017] (>2K citations)

Method	Mean IoU(%)	Pixel Acc.(%)
ResNet50-Baseline	37.23	78.01
ResNet50+B1+MAX	39.94	79.46
ResNet50+B1+AVE	40.07	79.52
ResNet50+B1236+MAX	40.18	79.45
ResNet50+B1236+AVE	41.07	79.97
ResNet50+B1236+MAX+DR	40.87	79.61
ResNet50+B1236+AVE+DR	**41.68**	**80.04**

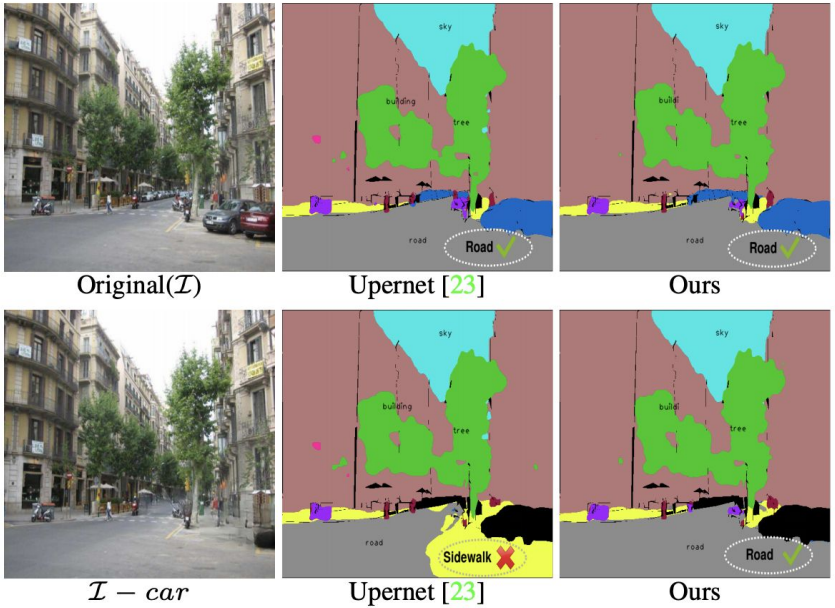
Table 1. Investigation of PSPNet with different settings. Baseline is ResNet50-based FCN with dilated network. 'B1' and 'B1236' denote pooled feature maps of bin sizes $\{1 \times 1\}$ and $\{1 \times 1, 2 \times 2, 3 \times 3, 6 \times 6\}$ respectively. 'MAX' and 'AVE' represent max pooling and average pooling operations individually. 'DR' means that dimension reduction is taken after pooling. The results are tested on the validation set with the single-scale input.

Semantic segmentation: modern approaches

- Pyramid Scene Parsing Network (PSPNet) [CVPR2017] (>2K citations)
  - CVPR 2017: http://openaccess.thecvf.com/content_cvpr_2017/papers/Zhao_Pyramid_Scene_Parsing_CVPR_2017_paper.pdf

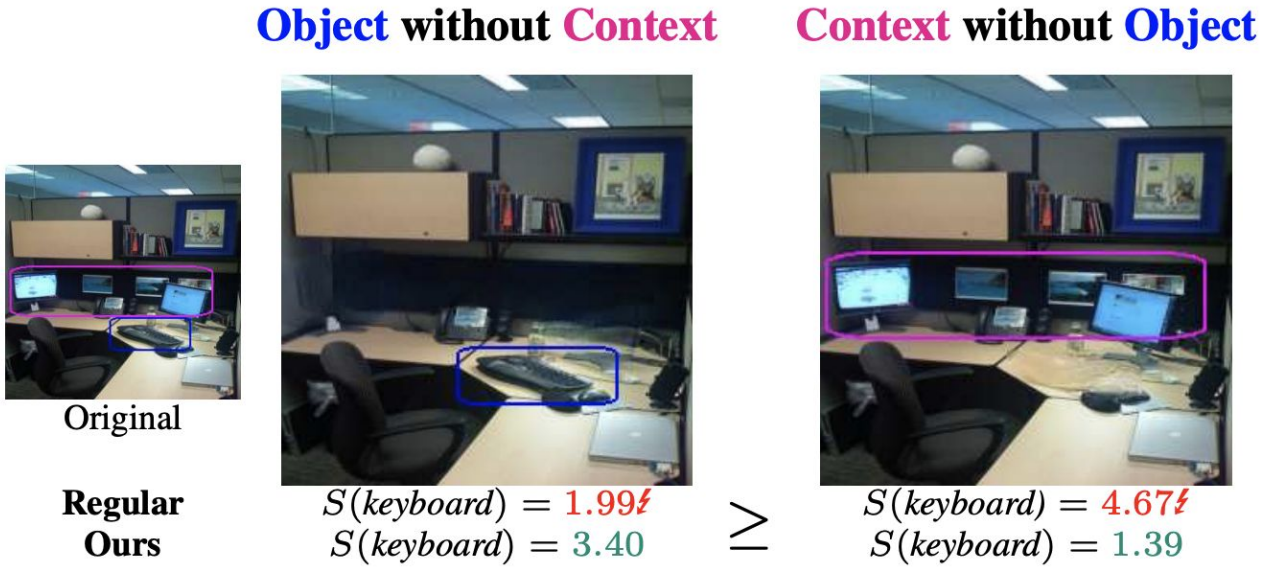
Semantic segmentation: modern approaches

- Not Using the Car to See the Sidewalk — Quantifying and Controlling the Effects of Context in Classification and Segmentation [CVPR2019]
  - Interesting paper about **influence of context**



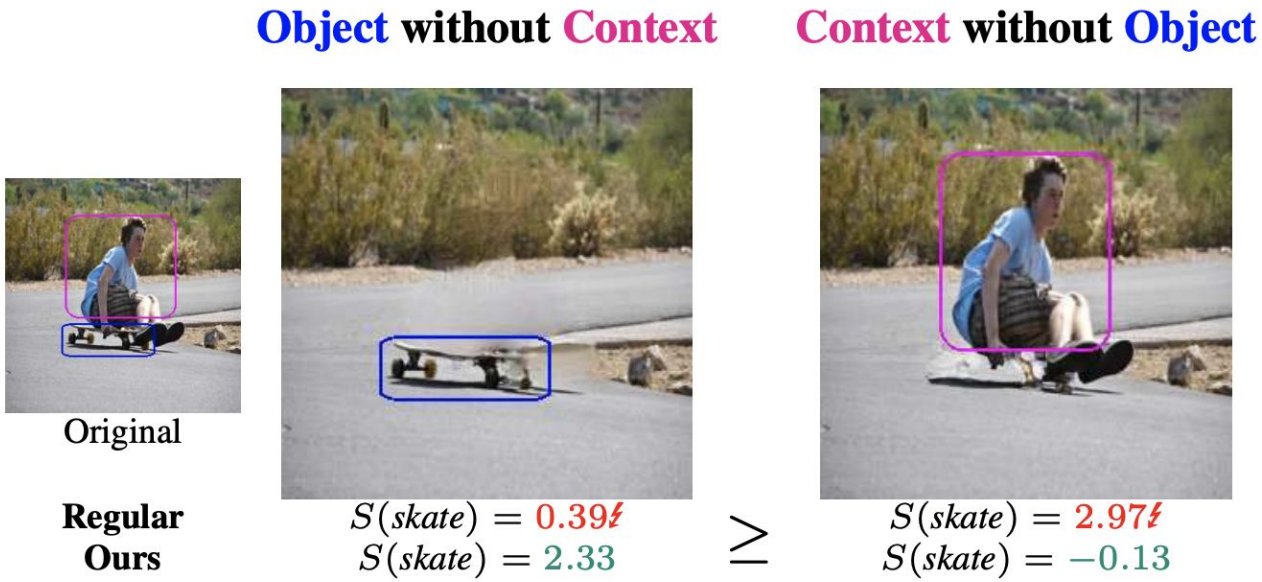
Semantic segmentation: modern approaches

- Not Using the Car to See the Sidewalk — Quantifying and Controlling the Effects of Context in Classification and Segmentation [CVPR2019]



Semantic segmentation: modern approaches

- Not Using the Car to See the Sidewalk — Quantifying and Controlling the Effects of Context in Classification and Segmentation [CVPR2019]



Semantic segmentation: modern approaches

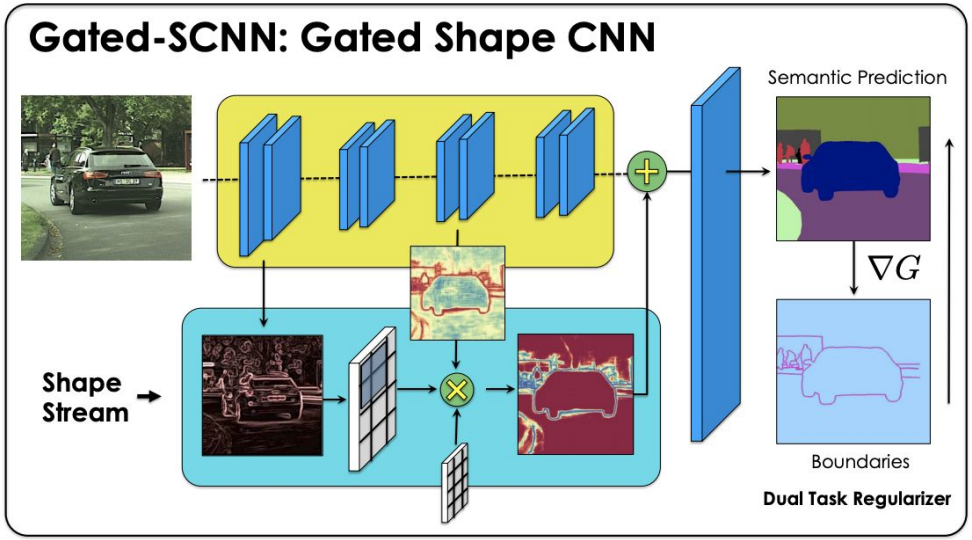
- Not Using the Car to See the Sidewalk — Quantifying and Controlling the Effects of Context in Classification and Segmentation [CVPR2019]
  - Object removal based data augmentation
    - Mitigate dependency between objects and context
    - Increase robustness of classification and segmentation models to contextual variations
  - Results obtained
    - Improve performance in out-of-context scenarios
    - Preserve performance on regular data

Semantic segmentation: modern approaches

- Not Using the Car to See the Sidewalk — Quantifying and Controlling the Effects of Context in Classification and Segmentation [CVPR2019]
  - http://openaccess.thecvf.com/content_CVPR_2019/papers/Shetty_Not_Using_the_Car_to_See_the_Sidewalk_--_Quantifying_CVPR_2019_paper.pdf

Semantic segmentation: modern approaches

- Gated-SCNN: Gated Shape CNNs for Semantic Segmentation [ICCV2019]
  - Motivation:
    - SoA methods process color, shape and texture information all together
    - Not ideal -> they may contain different type of information relevant for recognition
  - Idea: Two-stream CNN architecture
    - Classic stream
    - Shape stream

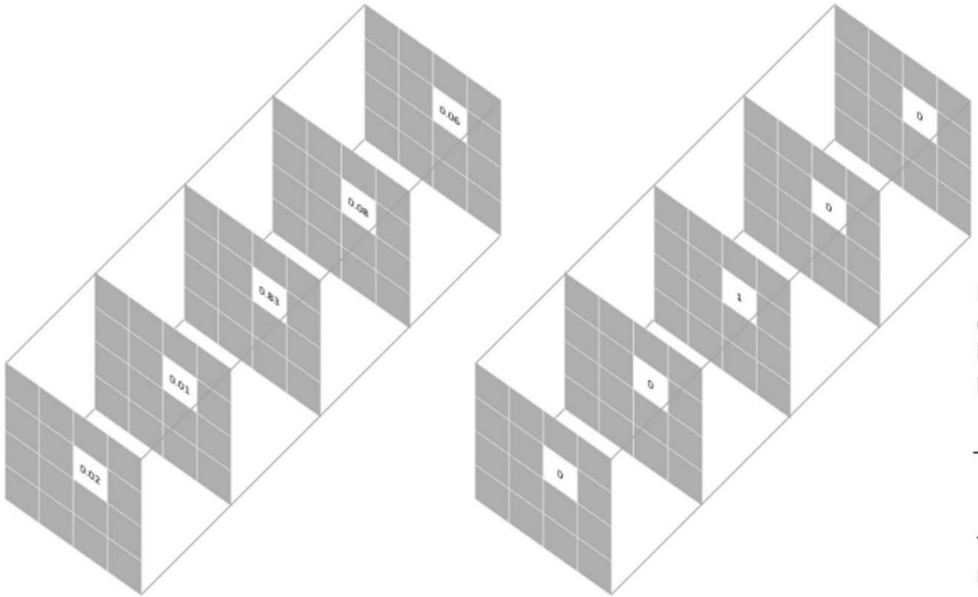


Semantic segmentation: modern approaches

- Gated-SCNN: Gated Shape CNNs for Semantic Segmentation [ICCV2019]
  - http://openaccess.thecvf.com/content_ICCV_2019/papers/Takikawa_Gated-SCNN_Gated_Shape_CNNs_for_Semantic_Segmentation_ICCV_2019_paper.pdf

Semantic segmentation: loss

- Pixel-wise cross entropy loss



Prediction for a selected pixel

Target for the corresponding pixel

Pixel-wise loss is calculated as the log loss, summed over all possible classes

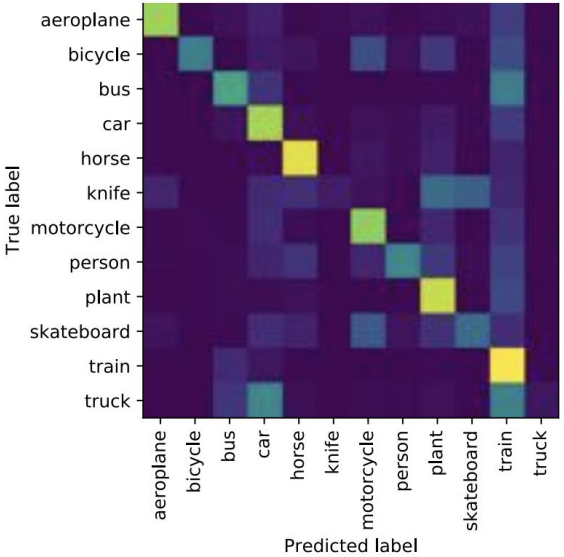
$$-\sum_{classes} y_{true} \log\left(y_{pred}\right)$$

This scoring is repeated over all **pixels** and averaged

Image source: https://www.jeremyjordan.me/semantic-segmentation/

Semantic segmentation: metrics

- Confusion matrix (same as in classification)
- Global accuracy
  - # correct pixels / total pixels
- Average per-class accuracy
  - avg. # correct pixels class(i) / total pixels class(i), for i=1...C
  - suitable for dataset with unlabeled categories
- Mean Intersection over Union (mIoU) (a.k.a. Jaccard)
  - avg. IoU(i) = TP / (TP + FP + FN) for all classes i=1...C
  - more strict than average per-class accuracy
  - penalizes false positive predictions
  - favors region smoothness and does not evaluate boundary accuracy



Semantic segmentation: datasets

- MS COCO
  - Common Objects in COntext
  - 80 categories
  - ~300K RGB images
  - Dense pixel-wise annotations for semantic segmentation (on superpixels)
  - Class and instance segmentation
  - http://cocodataset.org/#home

Semantic segmentation: datasets

- PASCAL Context
  - “everyday” images
  - 540 categories
  - ~10k RGB images (train)
  - Dense pixel-wise annotations for semantic segmentation
  - Class segmentation
  - https://cs.stanford.edu/~roozbeh/pascal-context/

Semantic segmentation: datasets

- ADE20k
  - Diverse set of scenes, objects and object parts
  - Large and unrestricted open vocabulary
  - ~20k RGB images (train)
  - ~2700 categories
  - Dense pixel-wise annotations for semantic segmentation
  - Class segmentation
  - https://groups.csail.mit.edu/vision/datasets/ADE20K/

Semantic segmentation: datasets

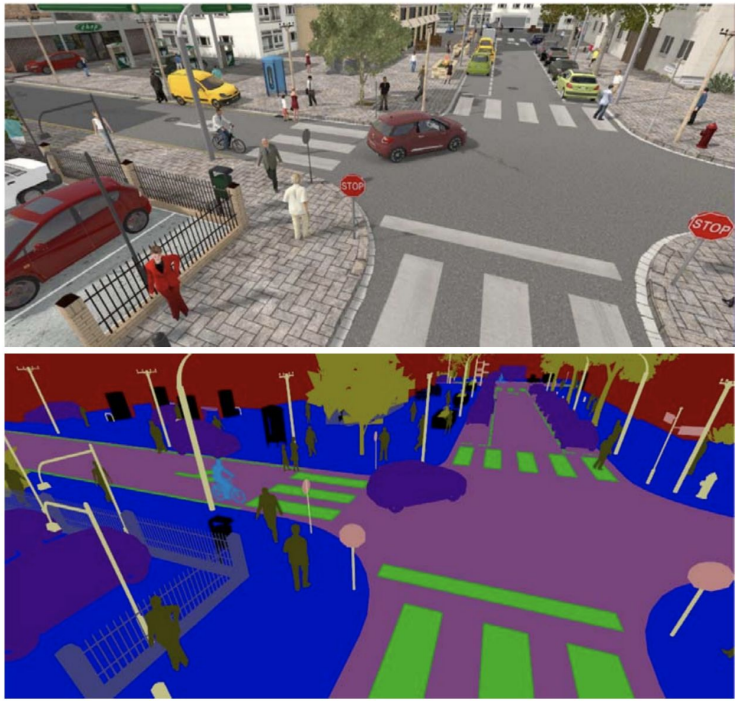
- Cityscapes
  - Urban driving scenarios
  - 30 categories
  - ~25k images with dense annotations
  - ~5k images with dense pixel-wise annotations for semantic segmentation
  - Class and instance segmentation
  - https://www.cityscapes-dataset.com/

Semantic segmentation: datasets

- Mapillary Vistas Dataset
  - Urban driving scenarios
  - 152 categories
  - ~25k images with dense pixel-wise annotations for semantic segmentation
  - Class and instance segmentation
  - Variety of weather, season, time of day, camera, and viewpoint
  - https://www.mapillary.com/dataset/vistas/

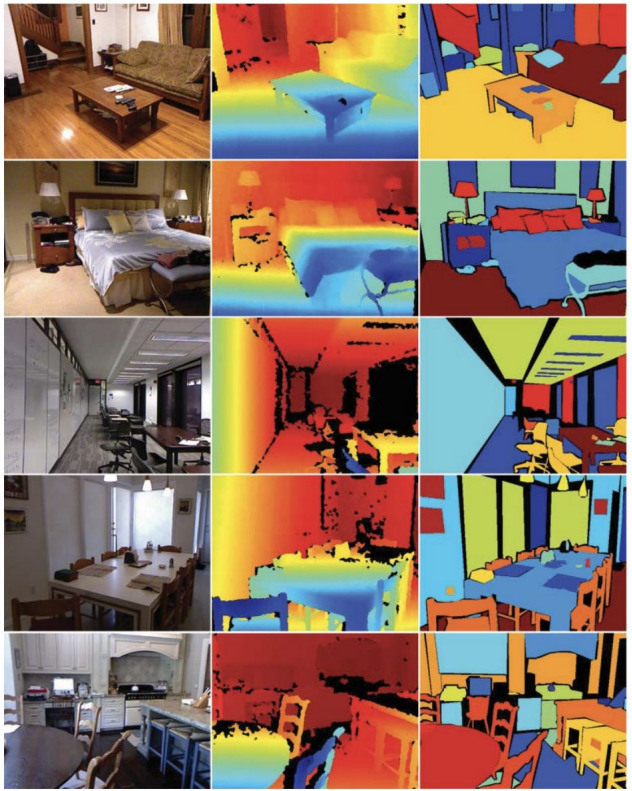
Semantic segmentation: datasets

- Synthia (from CVC!)
  - Synthetic driving scenarios
  - 13 categories
  - ~500k images with dense pixel-wise annotations for semantic segmentation
  - video streams
  - Class and instance segmentation
  - https://synthia-dataset.net/



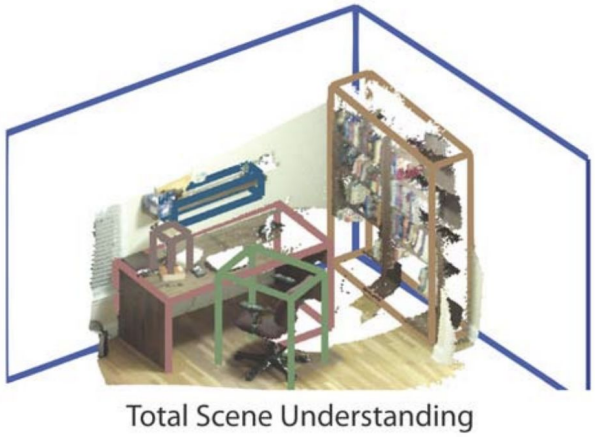
Semantic segmentation: datasets

- NYUD-V2
  - Real indoor scenes
  - 464 scenes
  - ~1500 RGB-D images
  - Dense pixel-wise annotations for semantic segmentation
  - Class and instance segmentation
  - https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html



Semantic segmentation: datasets

- SUN RGB-D
  - Real indoor scenes
  - ~10k RGB-D images
  - Dense pixel-wise annotations for semantic segmentation
  - Class and instance segmentation
  - http://rgbd.cs.princeton.edu/

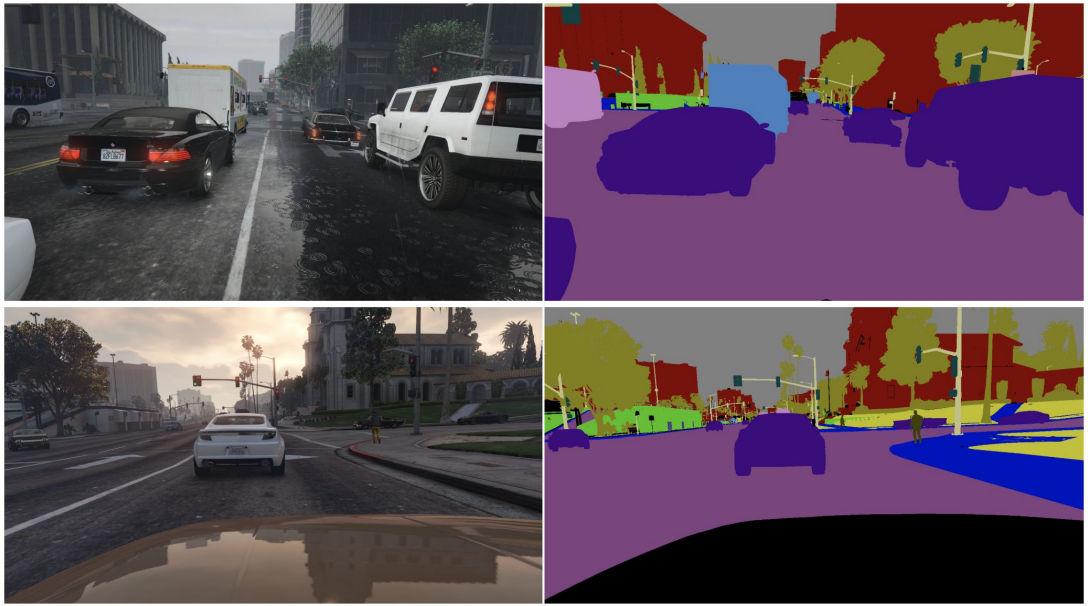


Semantic segmentation: datasets

- SceneNet RGB-D
  - Synthetic indoor scenes
  - ~5M RGB-D images
  - Dense pixel-wise annotations for semantic segmentation
  - Class and instance segmentation
  - https://robotvault.bitbucket.io/scenenet-rgbd.html

Semantic segmentation: datasets

- … and many more:
  - KITTI
  - Virtual KITTI
  - CamVid
  - SynthCity
  - LabelMe
  - GTA-V
  - Pascal Semantic Part
  - The Robotrix
  - ...

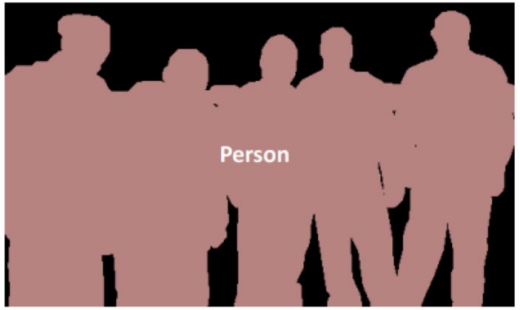


Outline

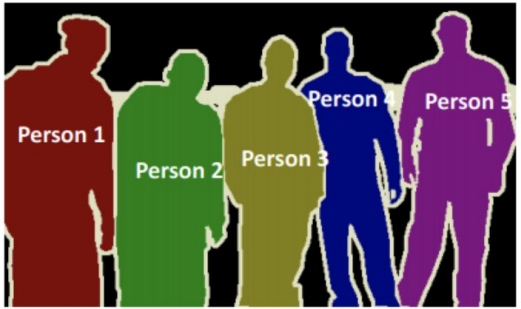
- Introduction to segmentation
- Semantic segmentation
- **Instance segmentation**
- Panoptic segmentation
- Amodal segmentation
- Referring image segmentation
- Current trends and future research

Instance segmentation: Problem statement

- Label regions that fully delineate object instances
- Analogous to object detection but with pixel-wise annotations
- More complete and thus harder than semantic segmentation
- Current state-of-the-art: object proposal + classification



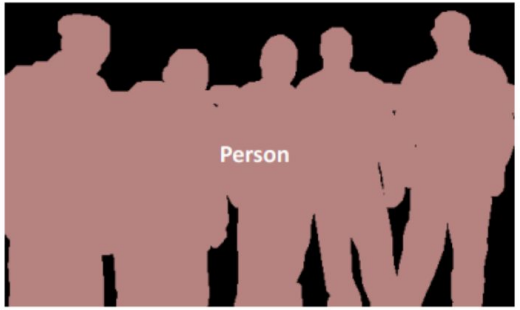
Semantic Segmentation



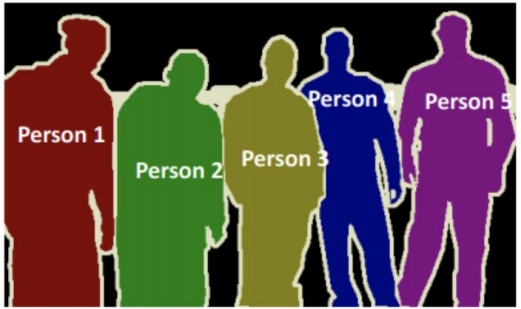
Instance Segmentation

Instance segmentation: Problem statement

- Harder than in semantic segmentation:
  - The representation of the problem is harder
  - It falls into the category of structure prediction
  - Naive representation does not work: do not try to use 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



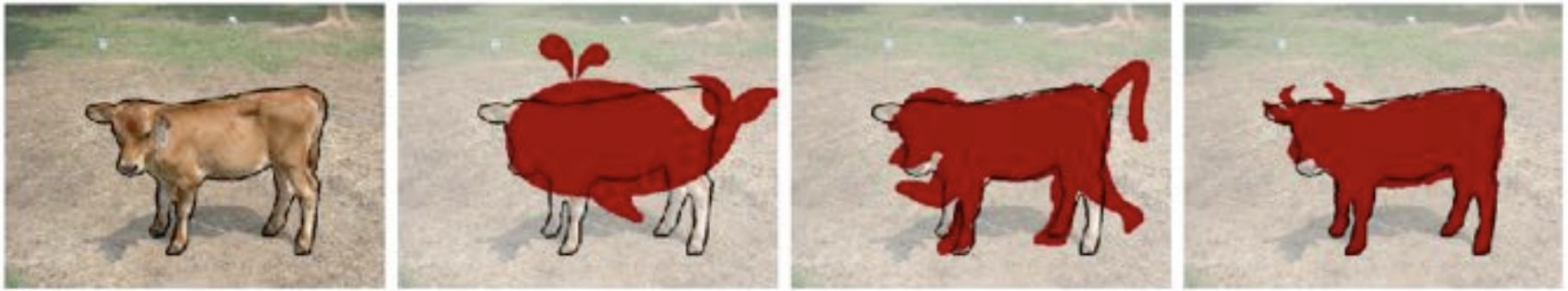
Semantic Segmentation



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

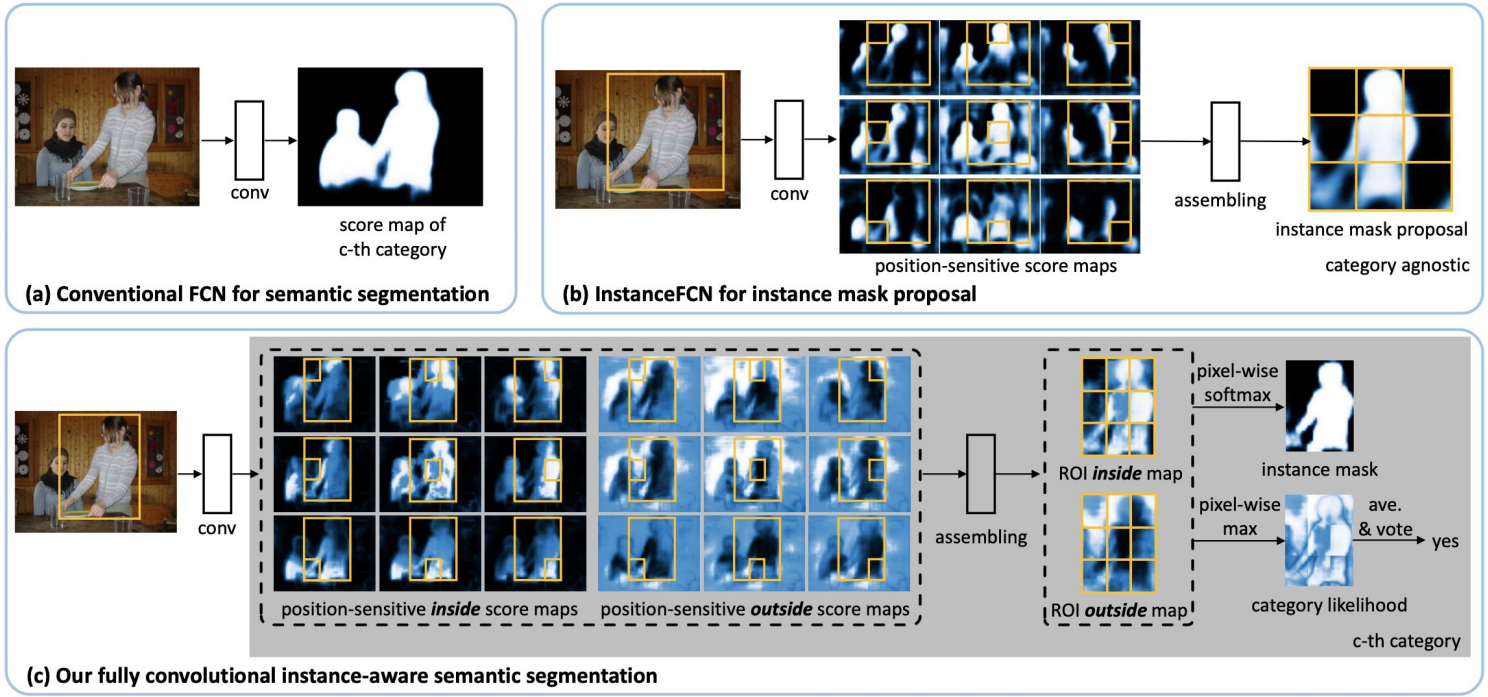
Figure from Krahenbuhl and Koltun. Geodesic Object Proposals. ECCV14

Instance segmentation: Region-based techniques

- Fully Convolutional Instance-aware Semantic Segmentation
  [CVPR2017] (~400 citations)
  - It detects and segments the object instances jointly and simultaneously
  - It combines:
    - FCNs for semantic segmentation
    - instance mask proposal

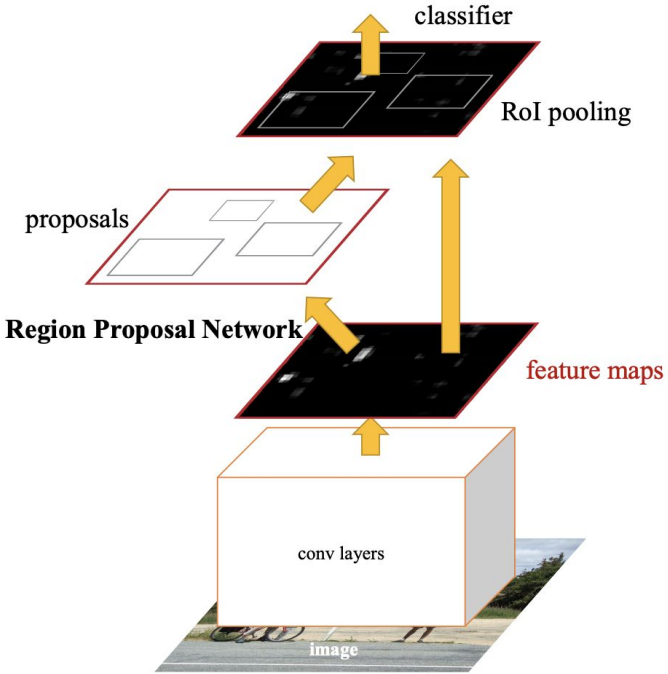
Instance segmentation: Region-based techniques

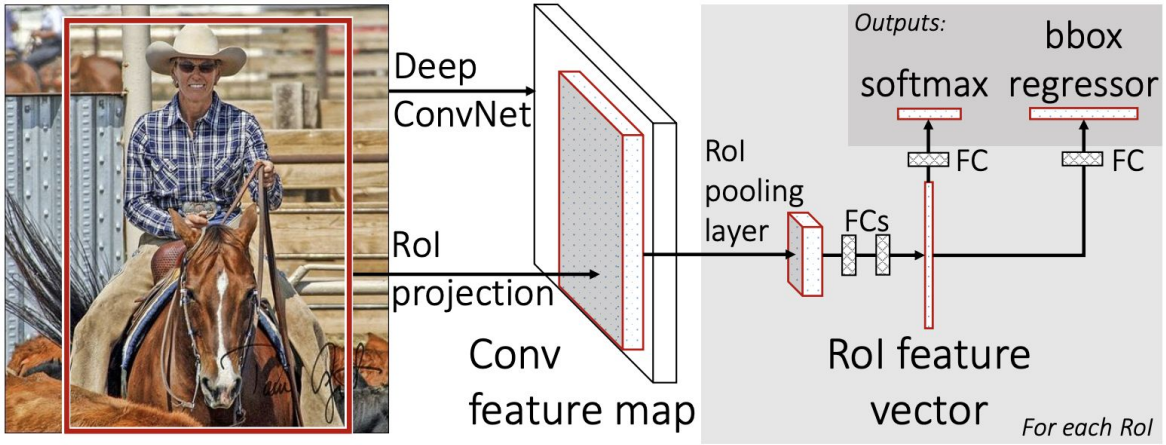
- Fully Convolutional Instance-aware Semantic Segmentation (FCIS) [CVPR2017]



Instance segmentation: Region-based techniques

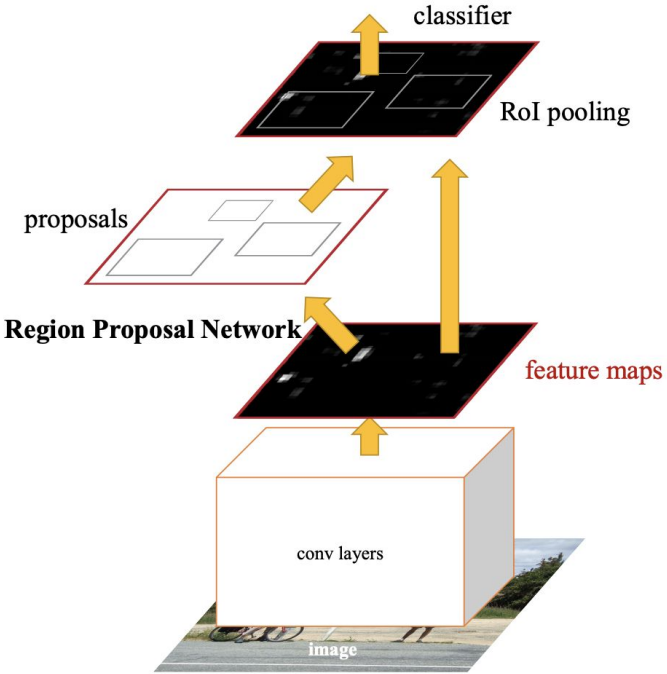
- Review SoA Object Detection: Fast R-CNN & Faster R-CNN





Instance segmentation: Region-based techniques

- Review SoA Object Detection: Fast R-CNN & Faster R-CNN



One network, four losses (TPAMI version):

- RPN classification (anchor good / bad)
- RPN regression (anchor -> proposal)
- Fast(er) R-CNN classification (over classes)
- Fast(er) R-CNN regression (proposal -> bbox)

Instance segmentation: Region-based techniques

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

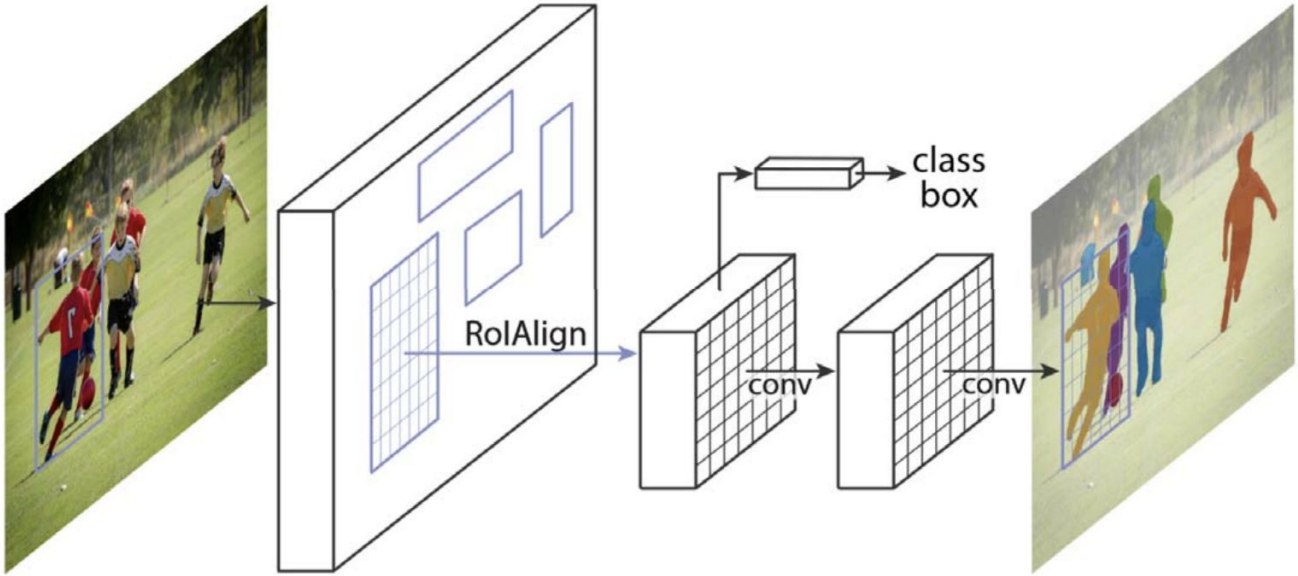


Image sourced from He, Kaiming, et al. "Mask R-CNN."

Instance segmentation: Region-based techniques

- Mask R-CNN [ICCV2017] (>5k citations)
  - Multi-task prediction:
    - object detection
    - mask
  - A region proposal network (RPN) proposes candidates
  - A mask prediction branch segments on such object proposals
  - It produces a segmentation mask for each category, but the segmentation output is the one with highest confidence in the “class” head

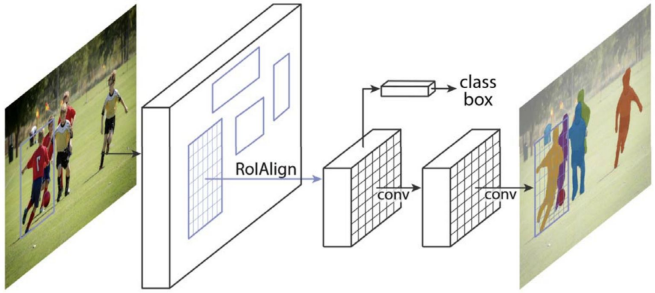
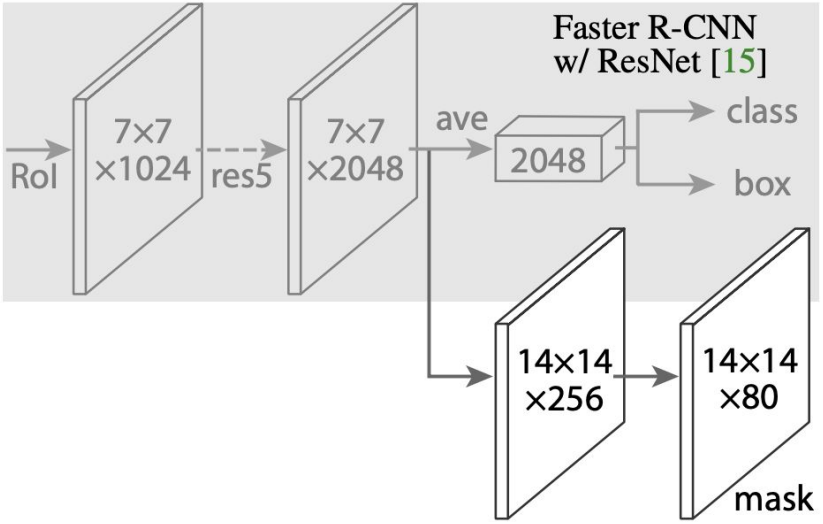


Image sourced from He, Kaiming, et al. "Mask R-CNN."

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 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

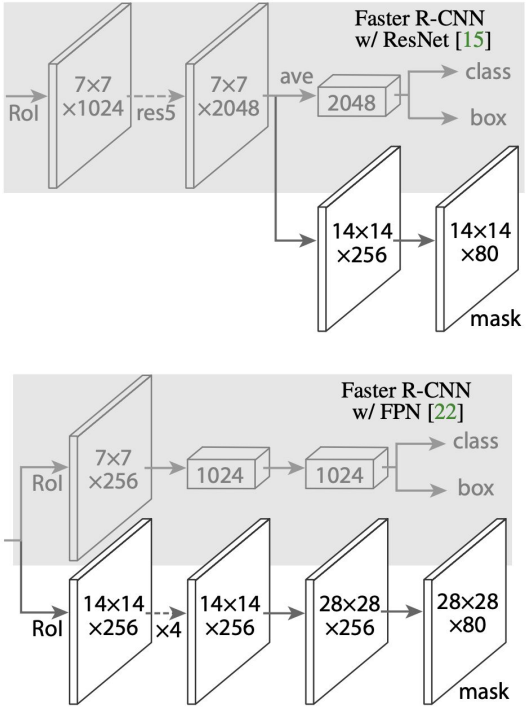
- Mask R-CNN [ICCV2017] (>5k citations)
  - Mask prediction requires estimation of pixel mask:
    - ROI pool (used in Fast(er) R-CNN): it contains 2 quantization steps -> ok for bounding box, but bad for pixel prediction
    - ROI pool -> ROI align: it uses bilinear interpolation to avoid quantization

	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

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: less than 10 layers from low-level structure to topmost features**
  - **Longpath: more than 100 layers from low-level structure to topmost features**

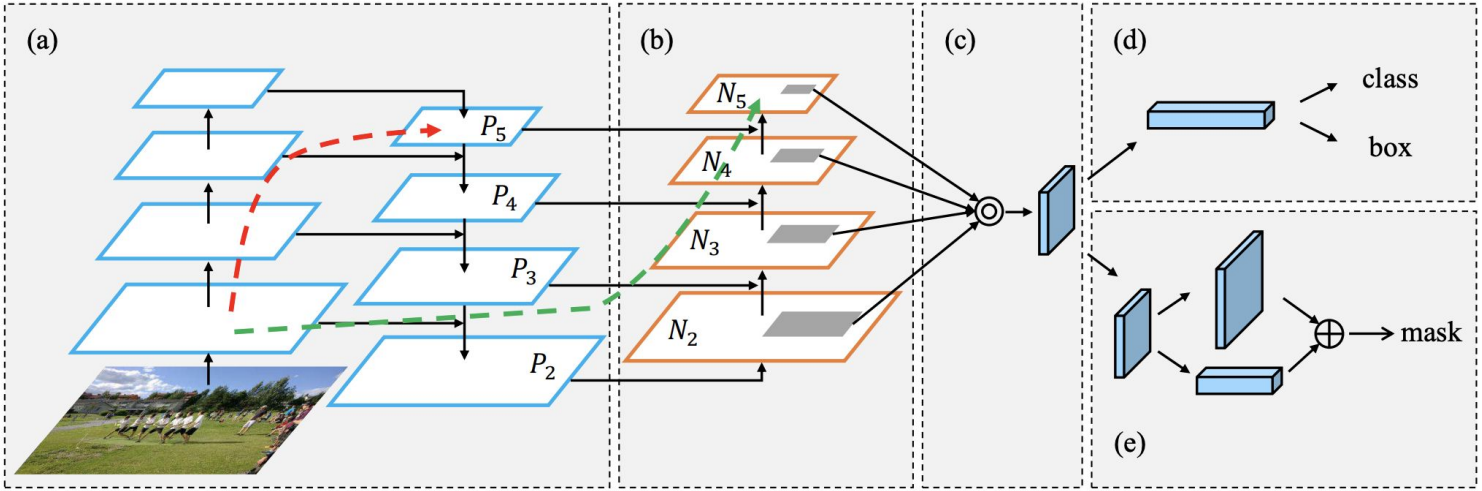


Figure 1. Illustration of our framework. (a) FPN backbone. (b) Bottom-up path augmentation. (c) Adaptive feature pooling. (d) Box branch. (e) Fully-connected fusion. Note that we omit channel dimension of feature maps in (a) and (b) for brevity.

Instance segmentation: Region-based techniques

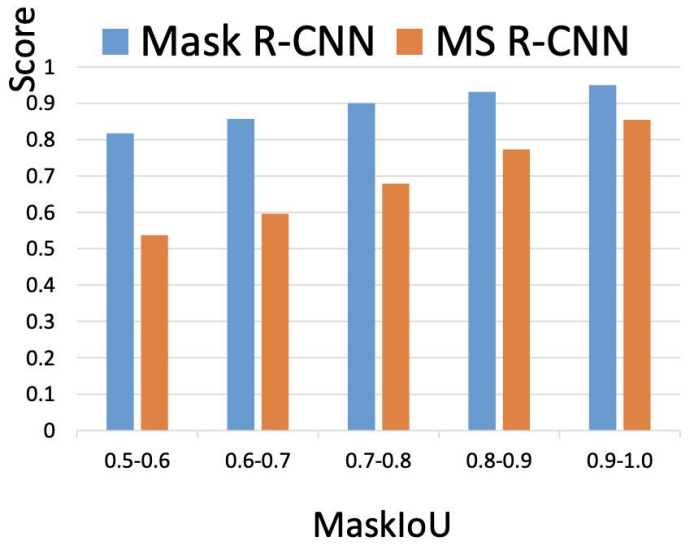
- Path Aggregation Network for Instance Segmentation (PA-Net) (>700 citations) [CVPR2018]
  - http://openaccess.thecvf.com/content_cvpr_2018/papers/Liu_Path_Aggregation_Network_CVPR_2018_paper.pdf

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:**
    - Including 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

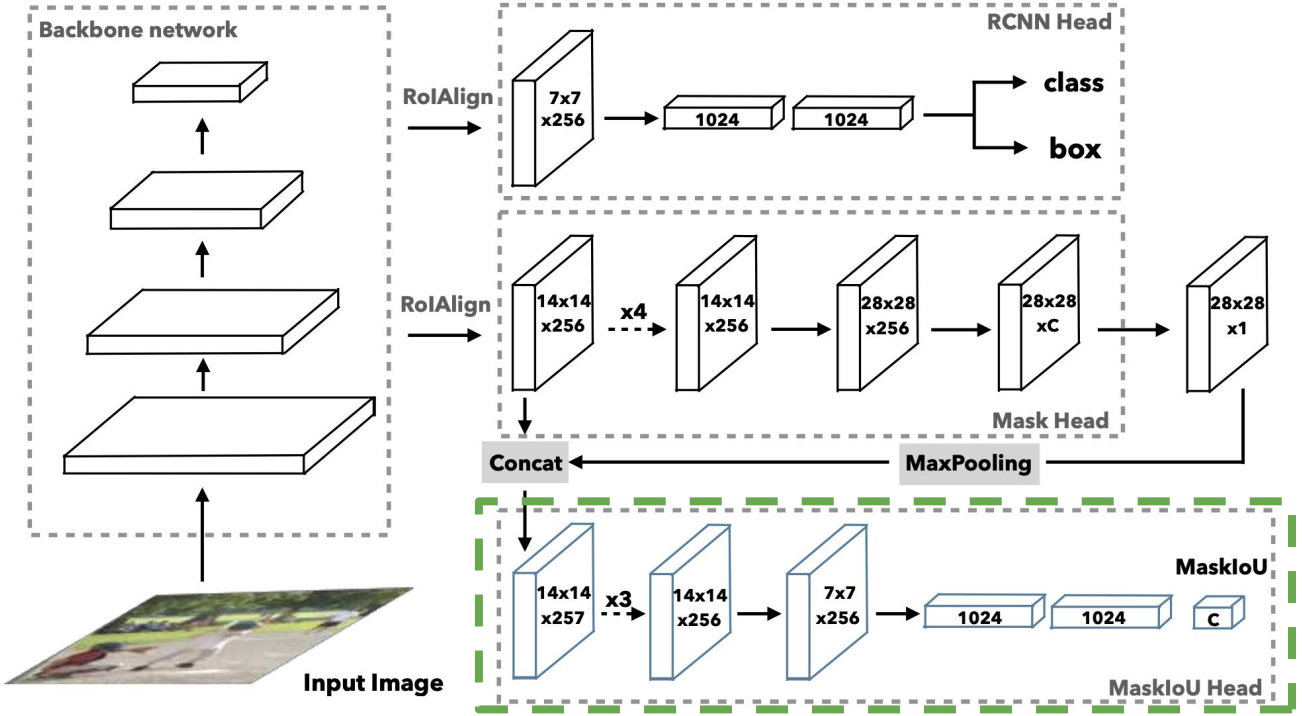
Instance segmentation: Region-based techniques

- 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

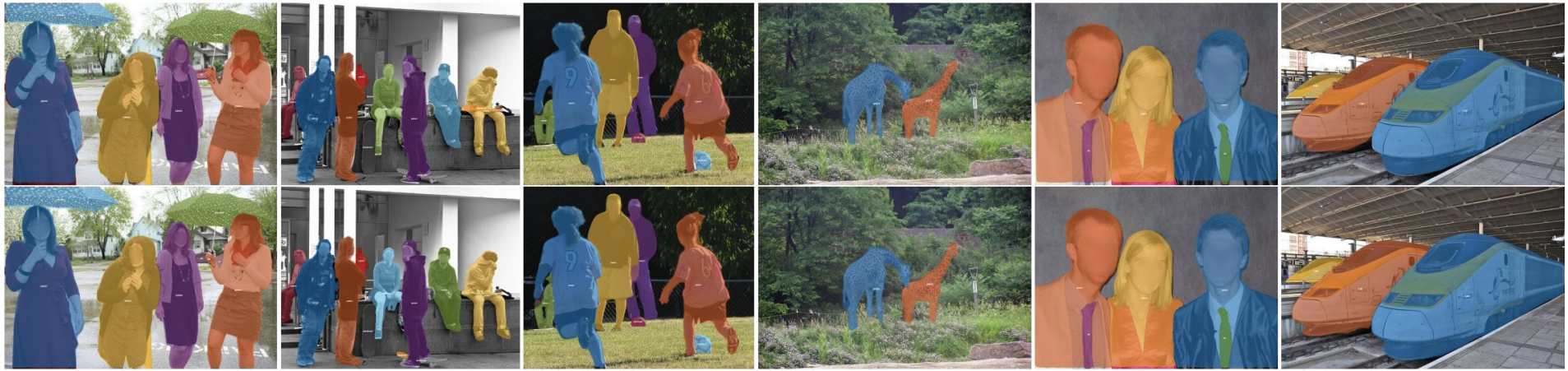
Instance segmentation: Region-based techniques

- TensorMask: A Foundation for Dense Object Segmentation [ICCV2019]
  - Same authors as Mask R-CNN (Facebook AI Research)
  - Instance segmentation approaches are dominated by methods that first detect object bounding boxes -> Mask R-CNN
  - TensorMask proposes **dense sliding window instance segmentation**
    - The output at every spatial location is itself a geometric structure with its own spatial dimensions
  - It achieves results close to the well-developed Mask R-CNN framework—both qualitatively and quantitatively.
    - It establishes a conceptually complementary direction for instance segmentation research.

Instance segmentation: Region-based techniques

- TensorMask: A Foundation for Dense Object Segmentation [ICCV2019]

Mask R-CNN



TensorMask

Instance segmentation: Region-based techniques

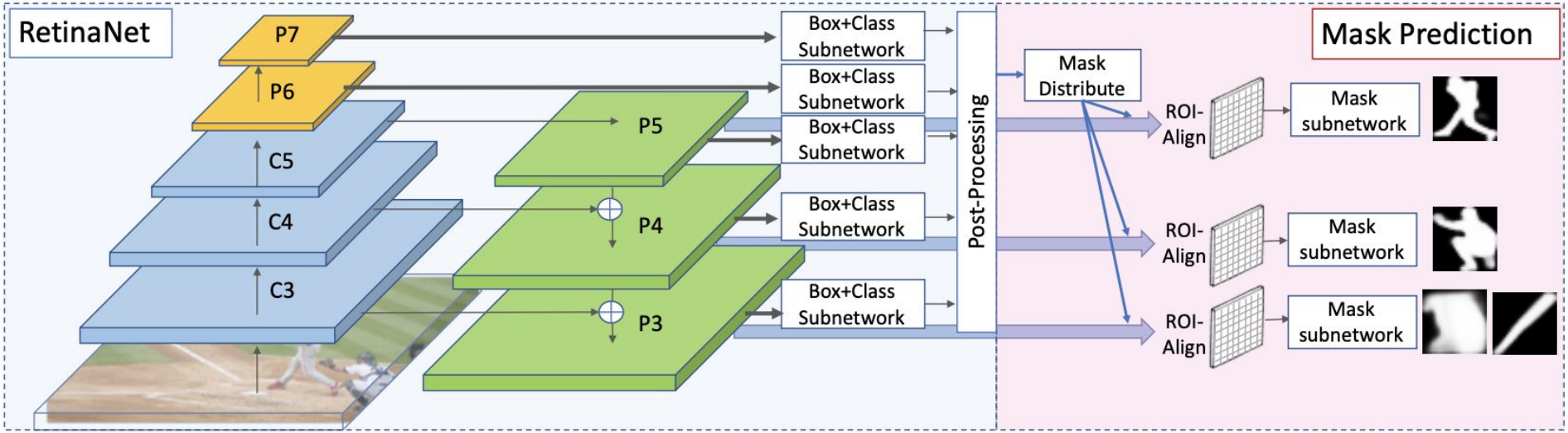
- TensorMask: A Foundation for Dense Object Segmentation [ICCV2019]
  - http://openaccess.thecvf.com/content_ICCV_2019/papers/Chen_TensorMask_A_Foundation_for_Dense_Object_Segmentation_ICCV_2019_paper.pdf

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
    - 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

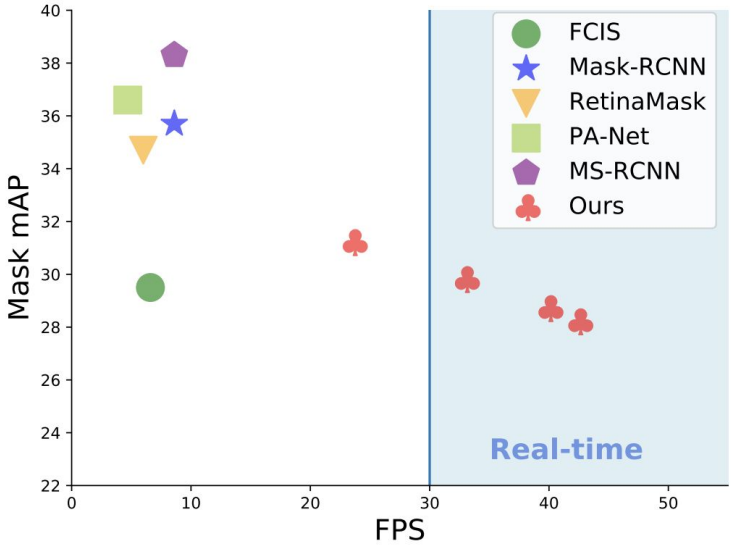


Instance segmentation: Region-based techniques

- RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free [arxiv2019]
  - https://arxiv.org/pdf/1901.03353.pdf

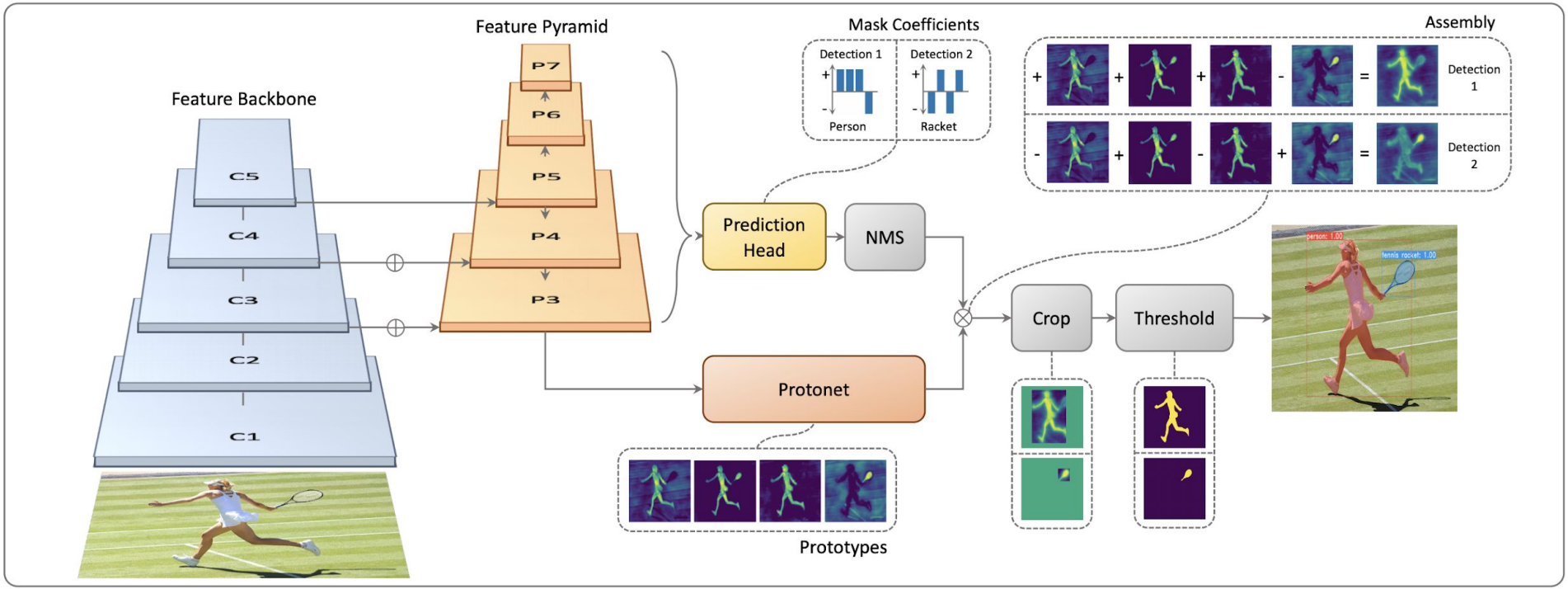
Instance segmentation: Region-based techniques

- 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

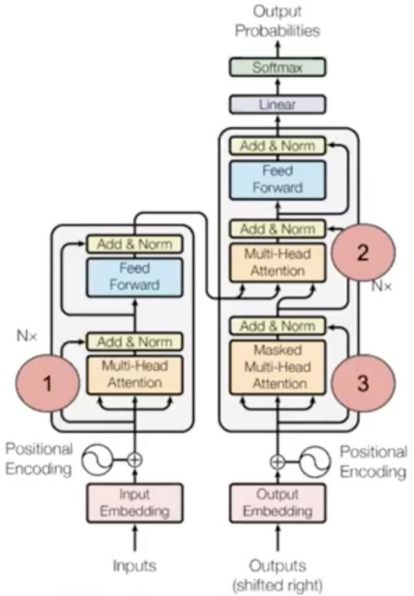
- YOLACT Real-time Instance Segmentation [ICCV2019]
  - http://openaccess.thecvf.com/content_ICCV_2019/papers/Bolya_YOLACT_Real-Time_Instance_Segmentation_ICCV_2019_paper.pdf

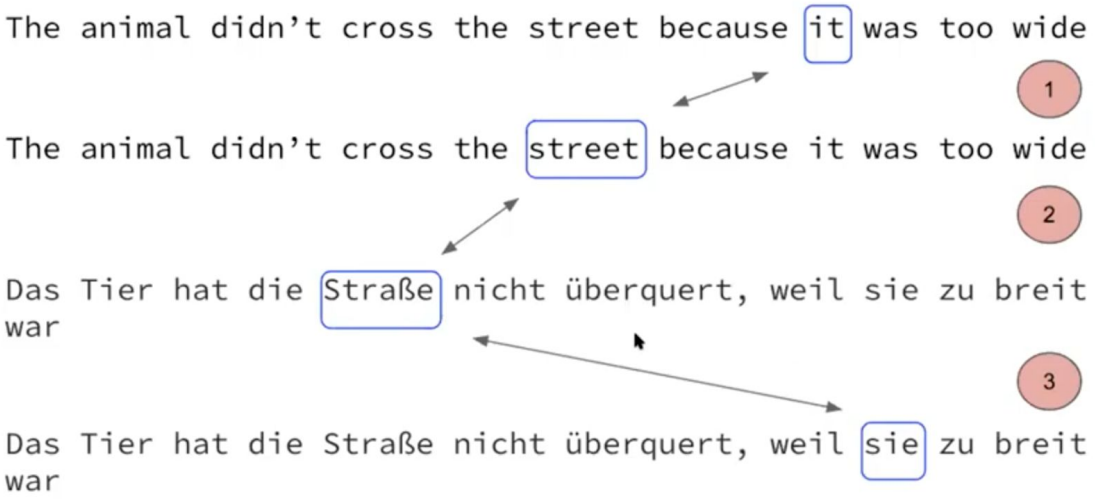
Instance segmentation: Region-based techniques

- End-to-end Object Detection with Transformers (DETR)
  - Approaches object detection as a direct set prediction problem.
  - It consists of a set-based global loss, which forces unique predictions via bipartite matching, and a Transformer encoder-decoder architecture.
  - 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

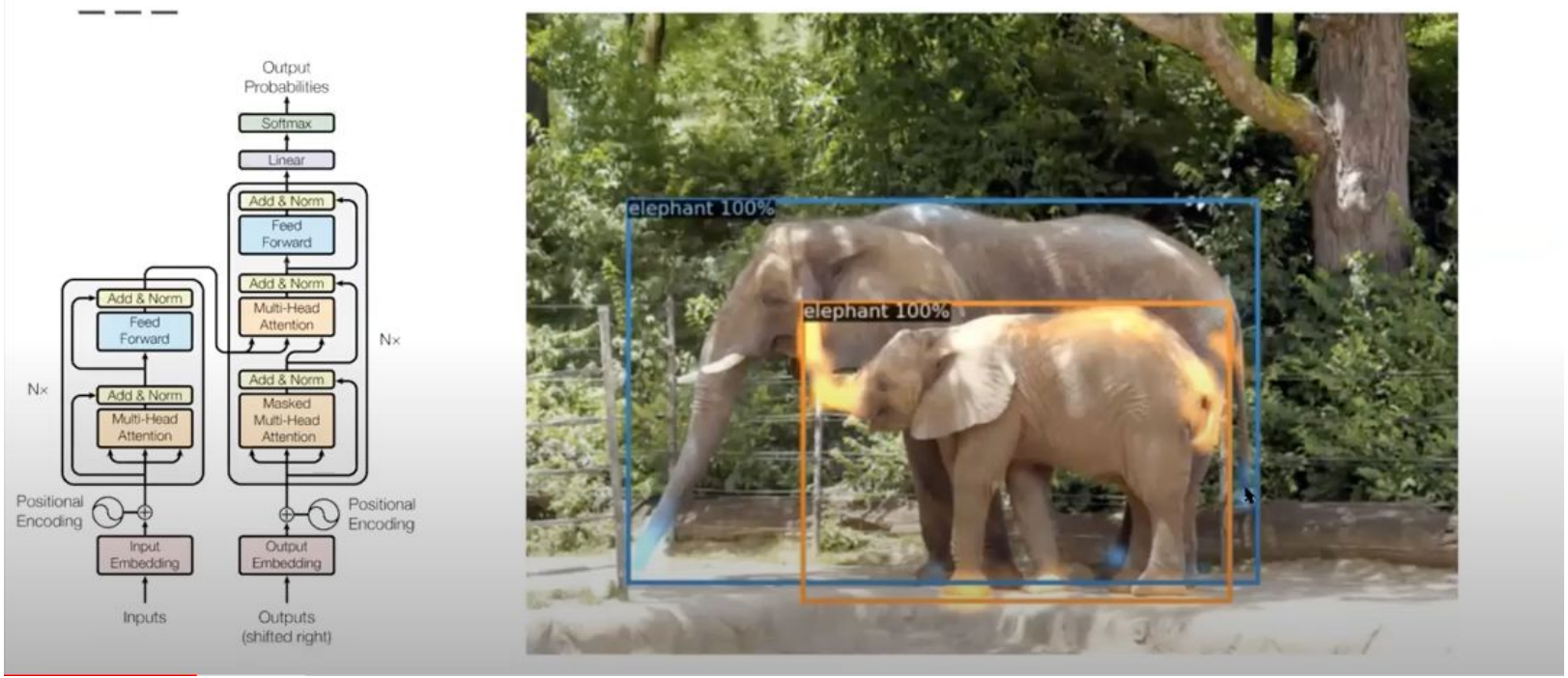
- End-to-end Object Detection with Transformers (DETR)





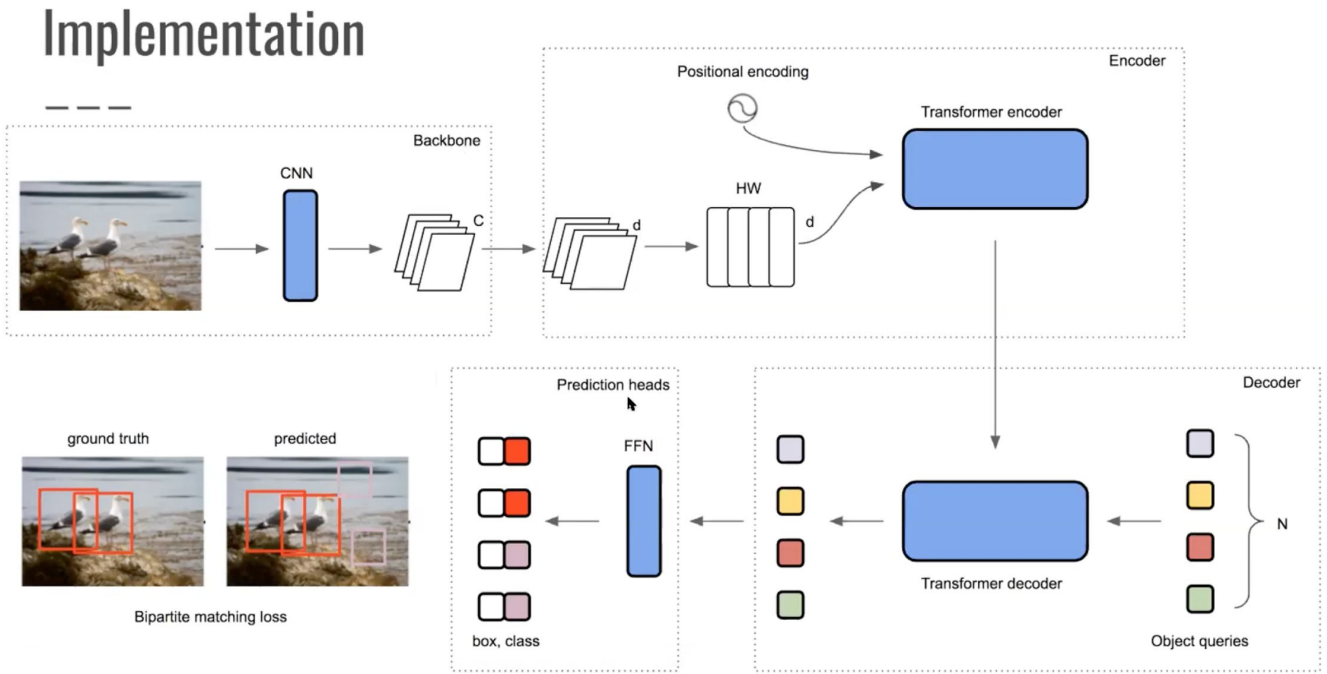
Instance segmentation: Region-based techniques

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



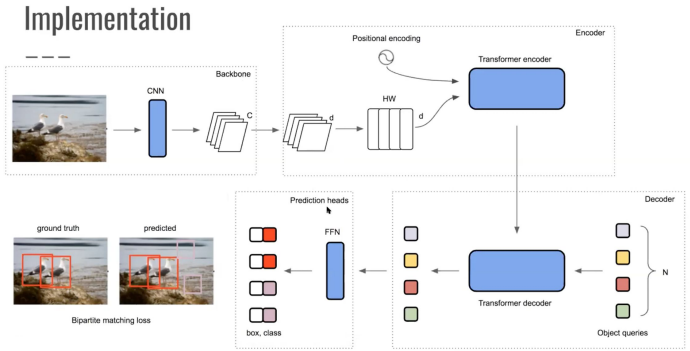
Instance segmentation: Region-based techniques

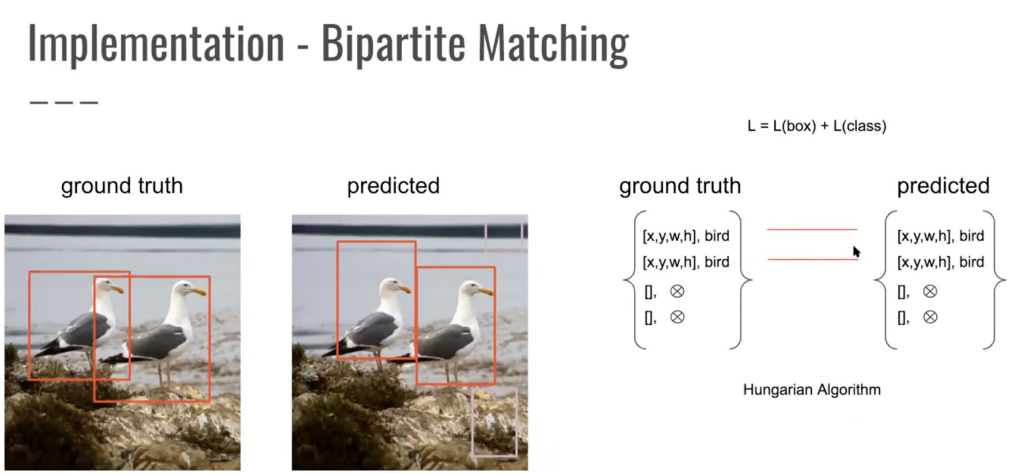
- End-to-end Object Detection with Transformers (DETR)



Instance segmentation: Region-based techniques

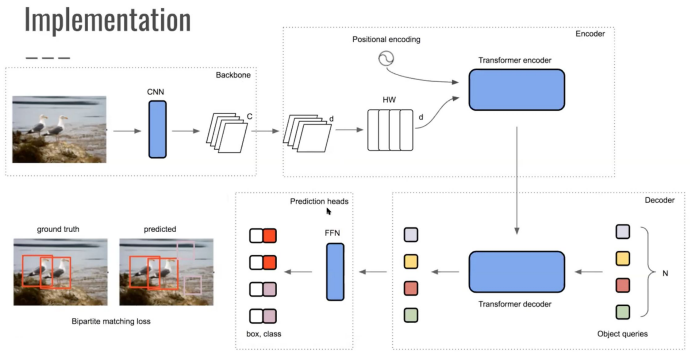
- End-to-end Object Detection with Transformers (DETR)

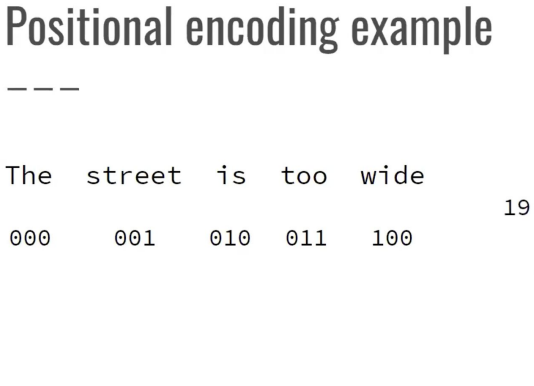




Instance segmentation: Region-based techniques

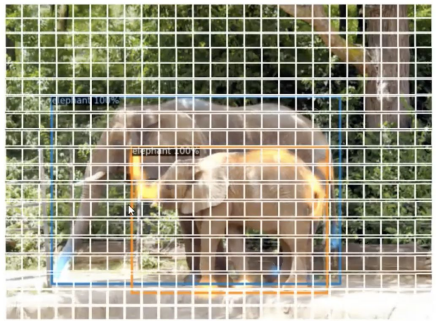
- End-to-end Object Detection with Transformers (DETR)





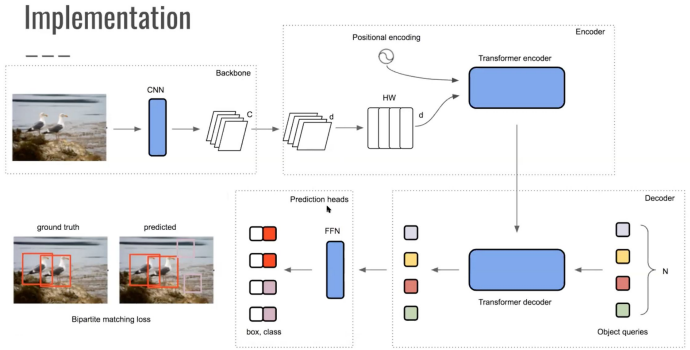
19

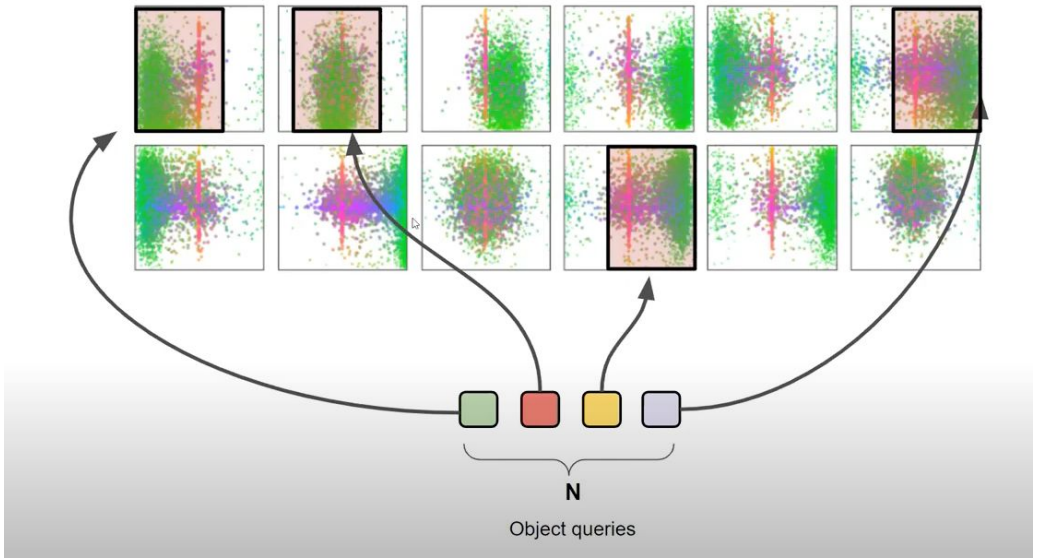
25



Instance segmentation: Region-based techniques

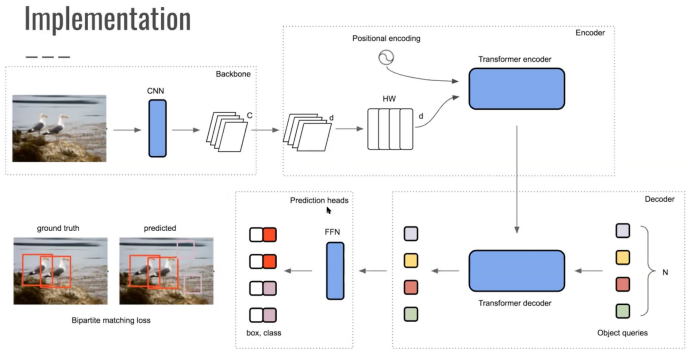
- End-to-end Object Detection with Transformers (DETR)

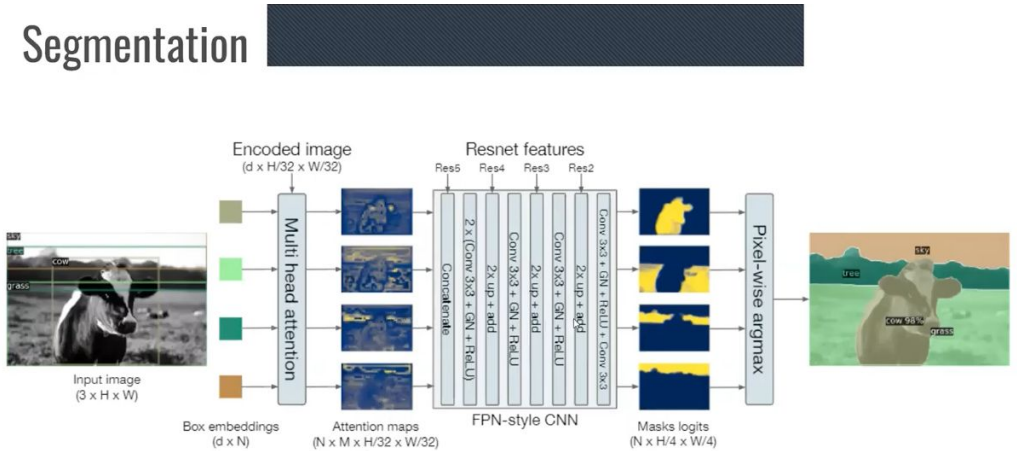




Instance segmentation: Region-based techniques

- 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)

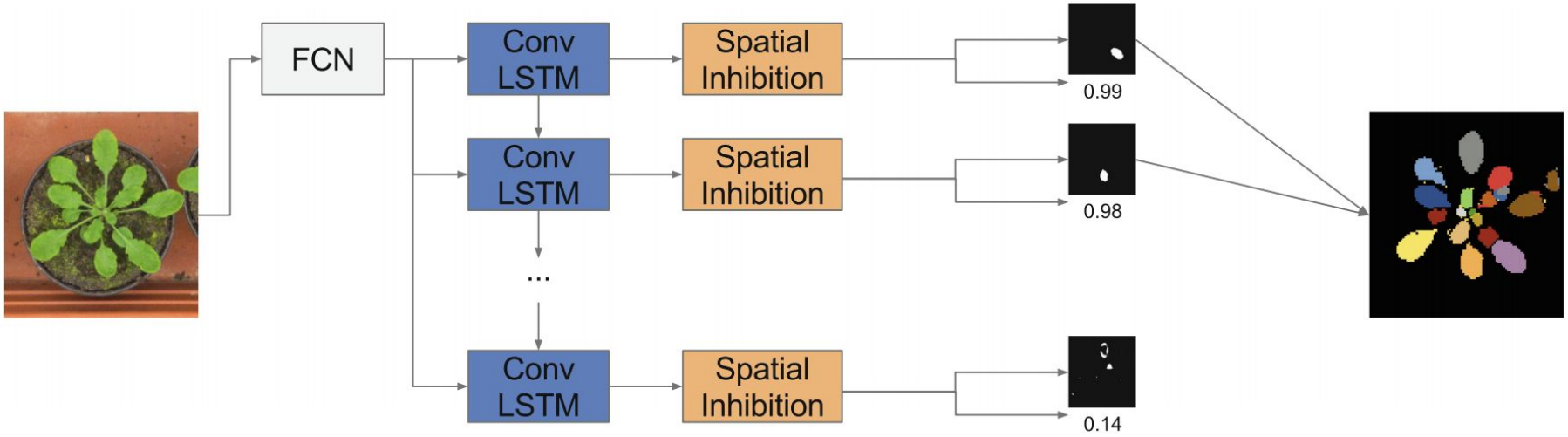


Fig. 1. Diagram of Recurrent Instance Segmentation (RIS).

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)

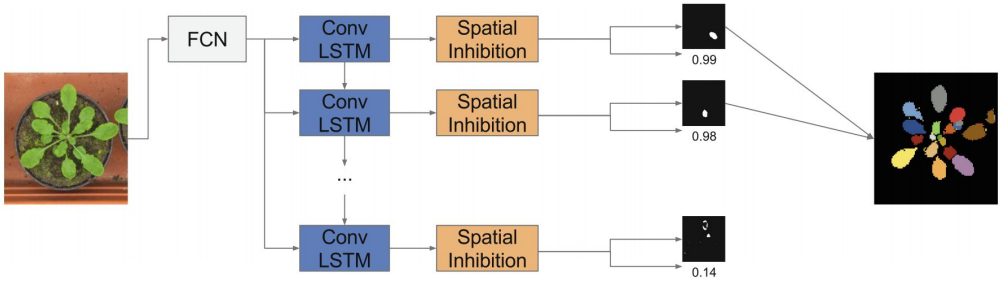


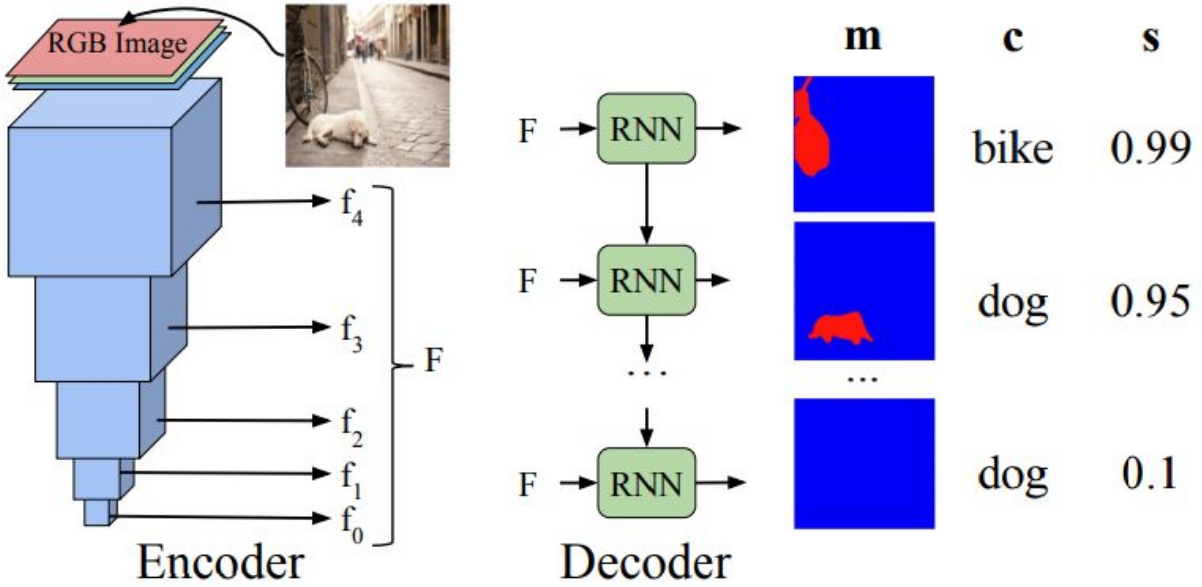
Fig. 1. Diagram of Recurrent Instance Segmentation (RIS).

Instance segmentation: RNN-based techniques

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

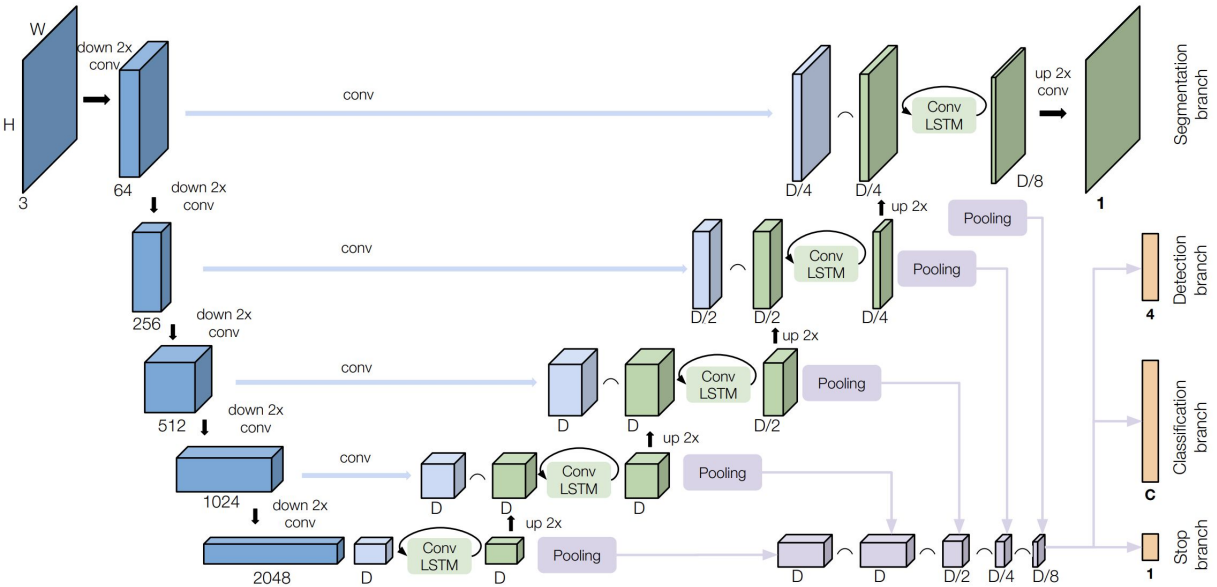
Instance segmentation: RNN-based techniques

- 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**

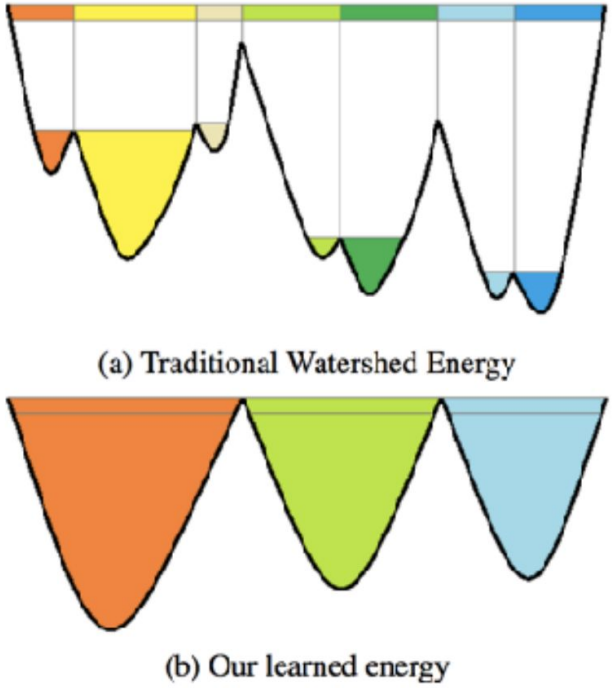
Table 1: Comparison against state of the art sequential methods for semantic instance segmentation. We specify whether the method is recurrent (Rec) and produces categorical probabilities (Cls).

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 given by semantic segmentation map

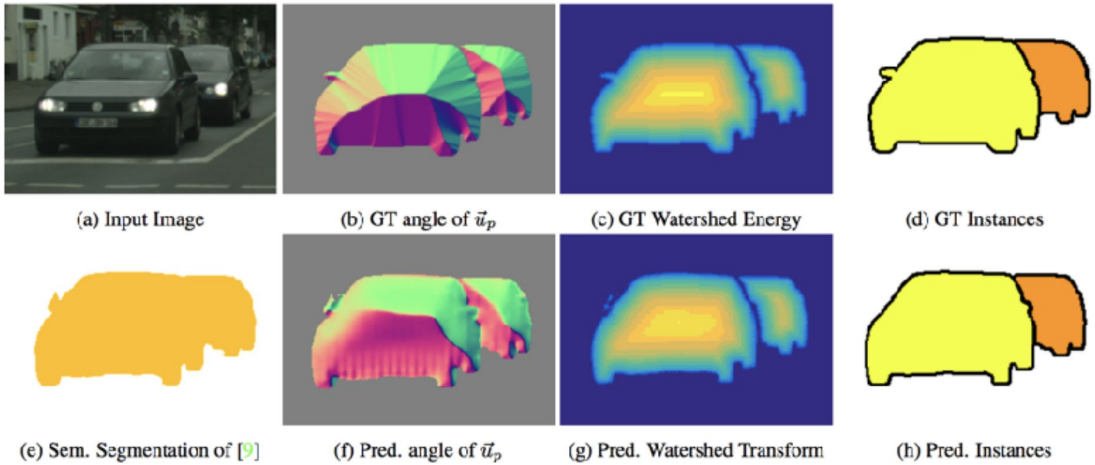


(a) Traditional Watershed Energy

(b) Our learned energy

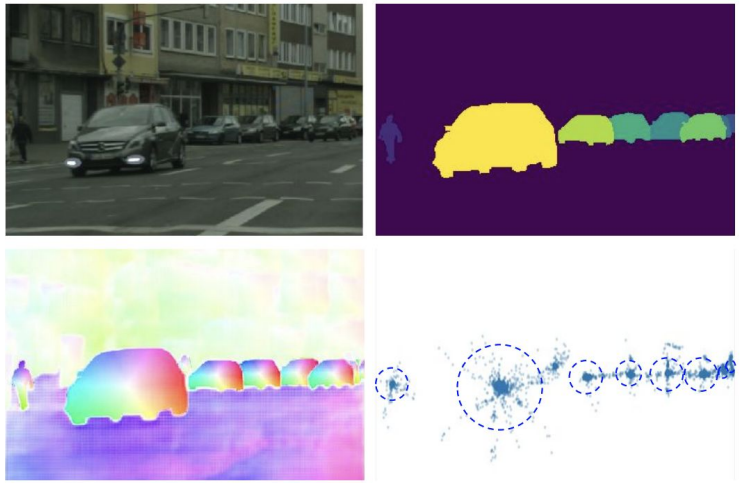
Instance segmentation: Partition space techniques

- 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 pixel with wrong object -> **maximum angular penalty**
    - **Task 2:** Predict energy function



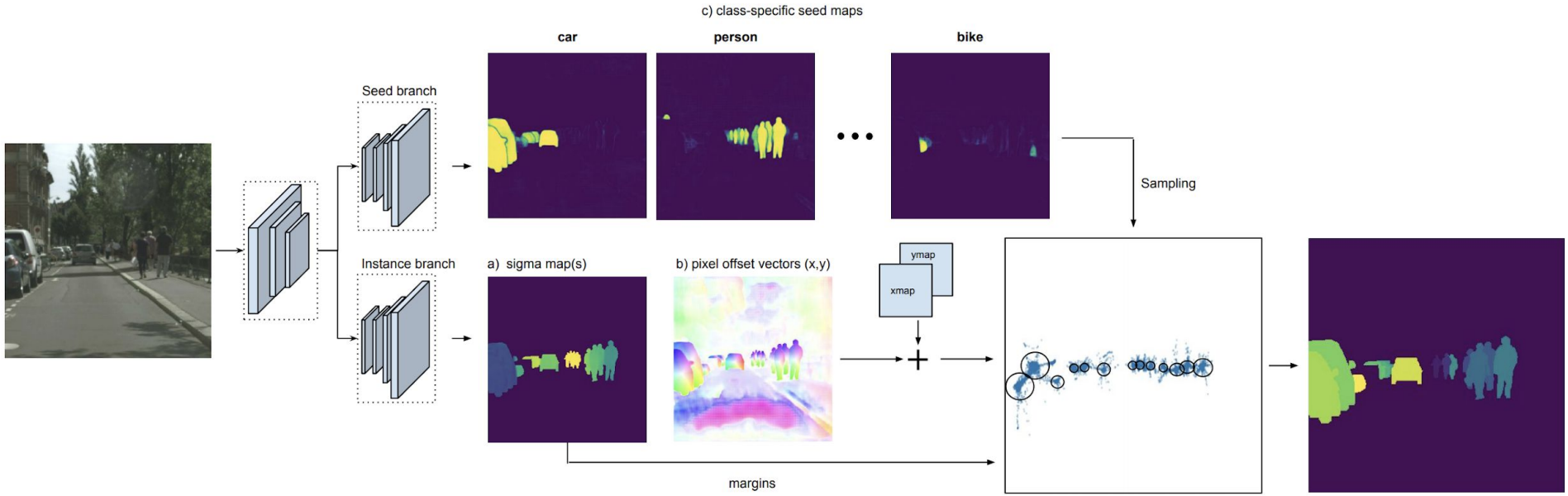
Instance segmentation: Partition space techniques

- 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)

Instance segmentation: datasets

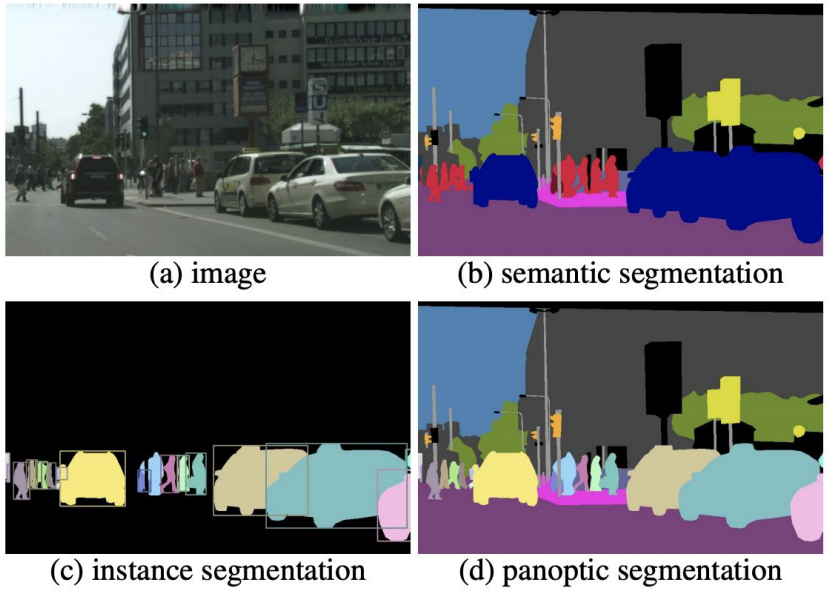
- New datasets have appeared:
  - LVIS: A Dataset for Large Vocabulary Instance Segmentation [CVPR2019]

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)
  - Instance segmentation (detect and segment each object instance)



(a) image (b) semantic segmentation

(c) instance segmentation (d) panoptic segmentation

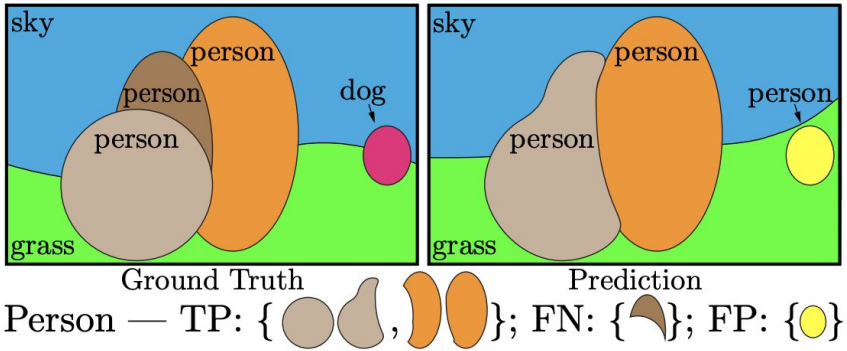
Panoptic segmentation: problem statement

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

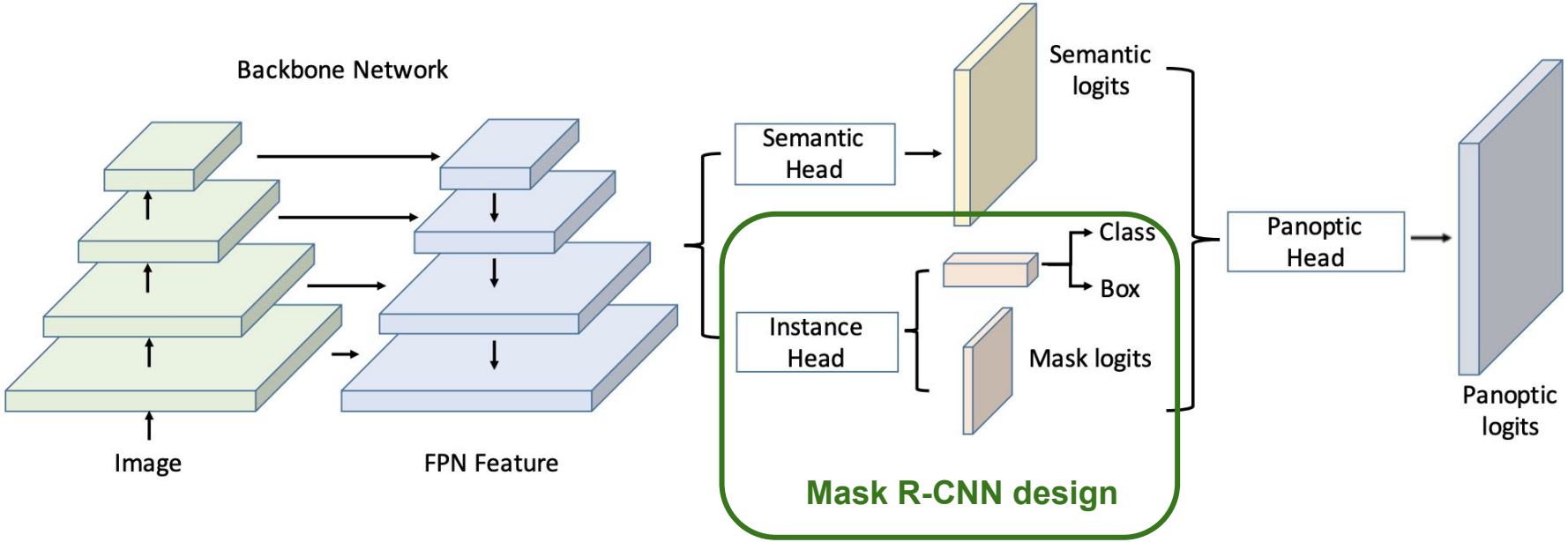
Panoptic segmentation: datasets

- MS COCO
- Cityscapes
- ADE20k
- Mapillary Vistas

COCO 2019 Panoptic Segmentation Task

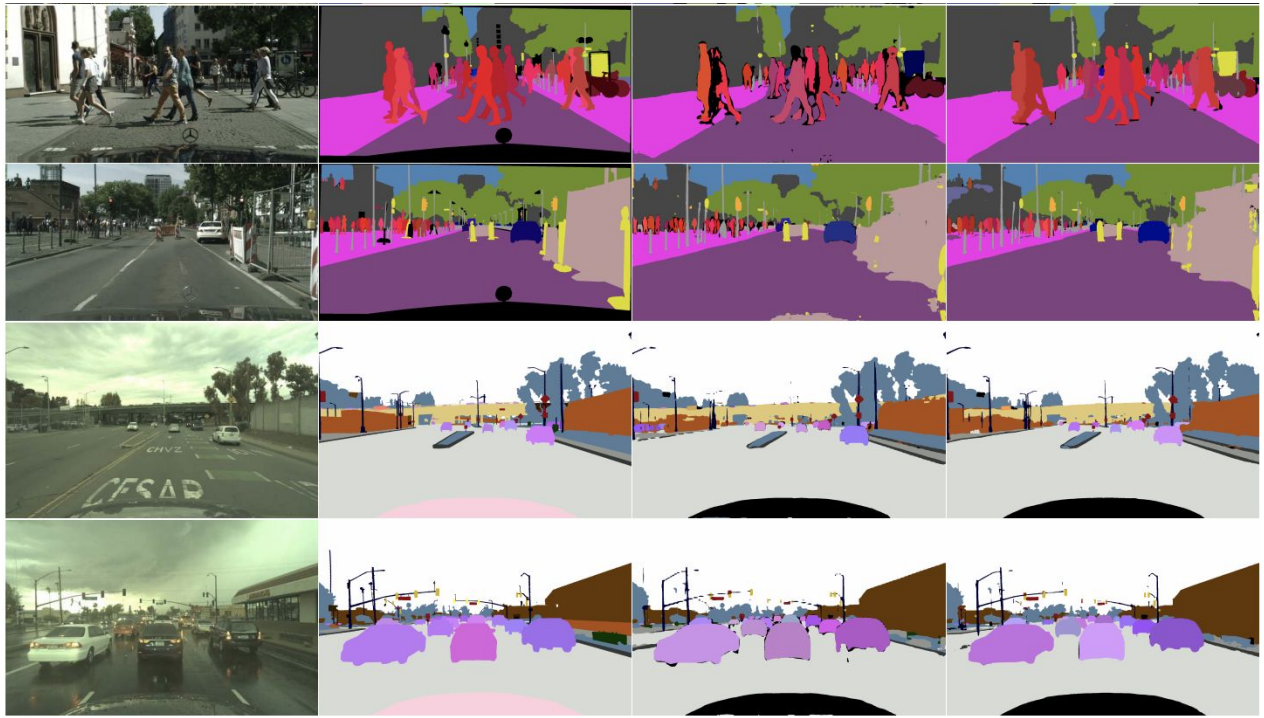
Panoptic segmentation: techniques

- 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

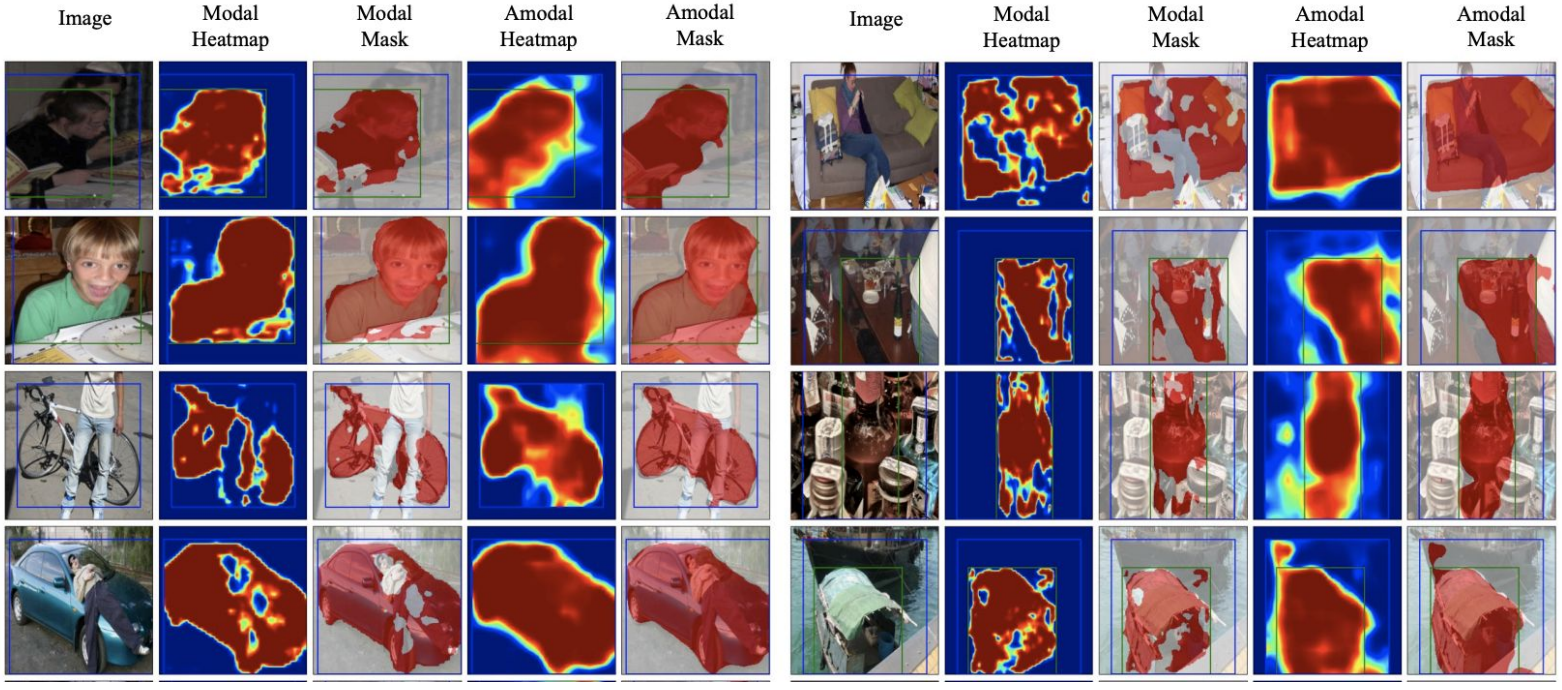
- Objective: predict the region encompassing both **visible and occluded** parts of each object.
- Problem defined in [ECCV2016]

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

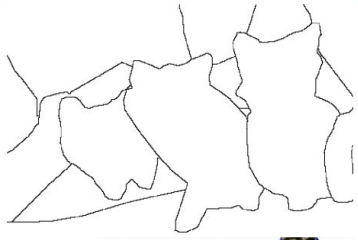
Amodal segmentation: techniques

- Semantic Amodal Segmentation [CVPR2017]

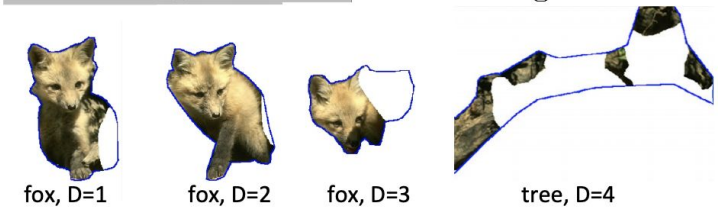
input image

modal segmentation

depth ordering



edge annotations



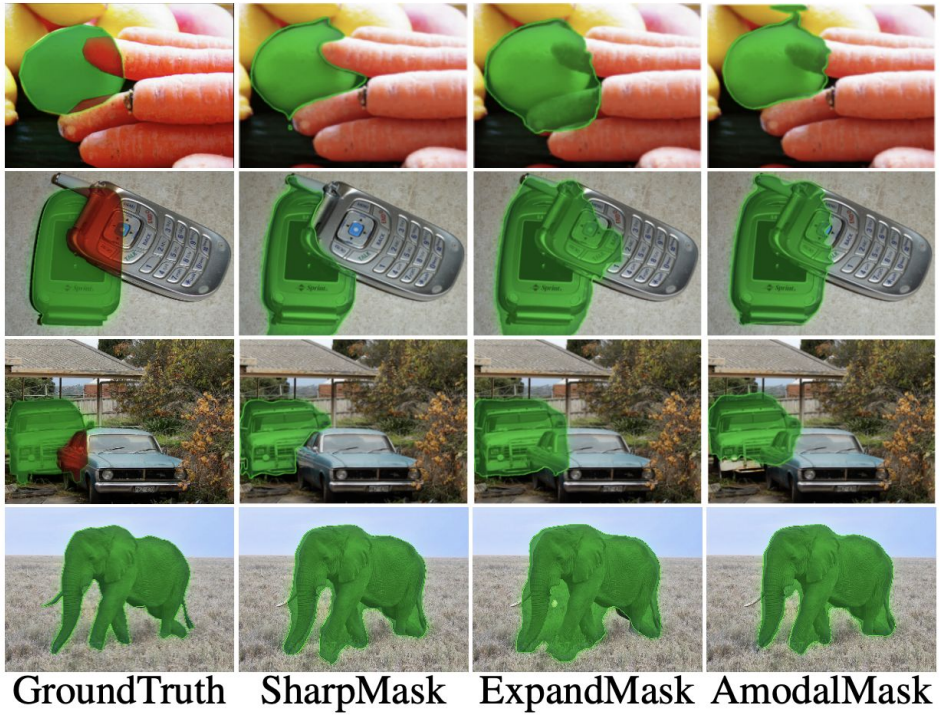
amodal segmentation
+
semantic + depth order

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 modal segmentation but computing 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

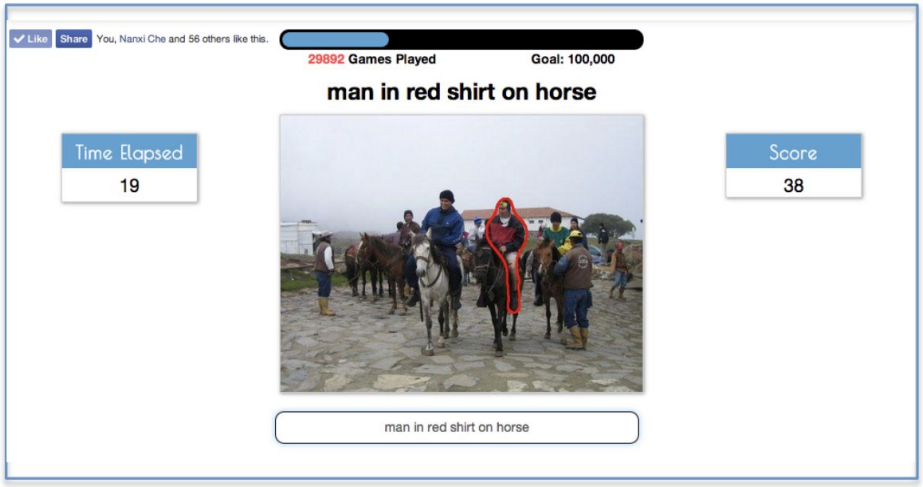
- The task of referring expression comprehension is to localize a region described by a given referring expression

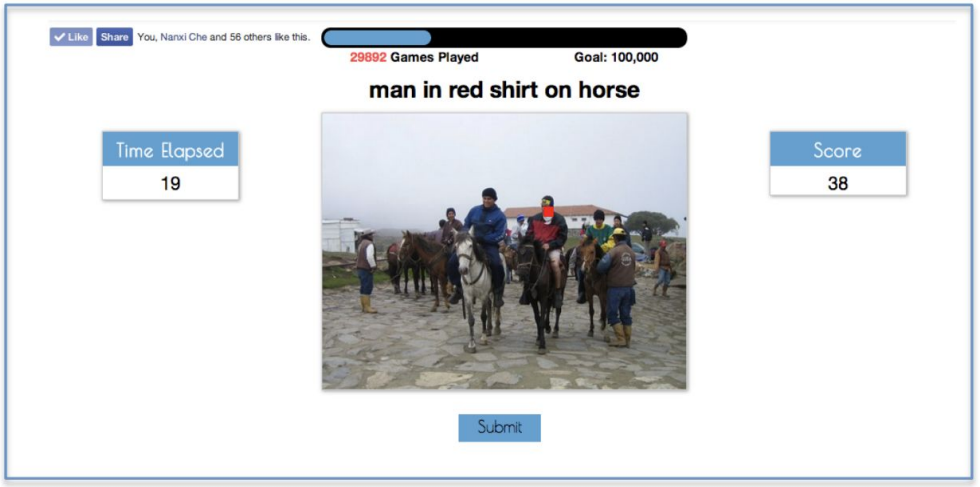
Referring image segmentation: problem definition

- Assumption about referring expressions:
  - A referring expression can’t be ambiguous, it should identify the object from the scene without any ambiguity
- 4 principles about natural language dialogue interactions [ECCV2016]:
  - 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]



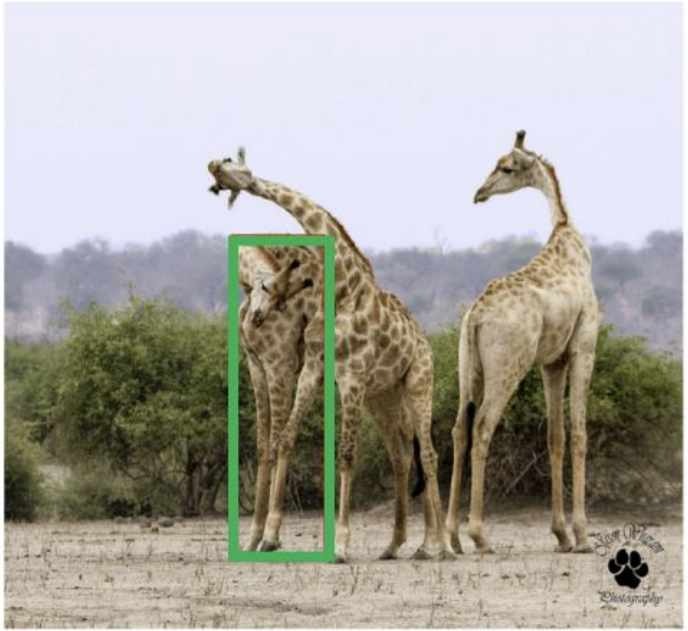


Referring image segmentation: techniques

- 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

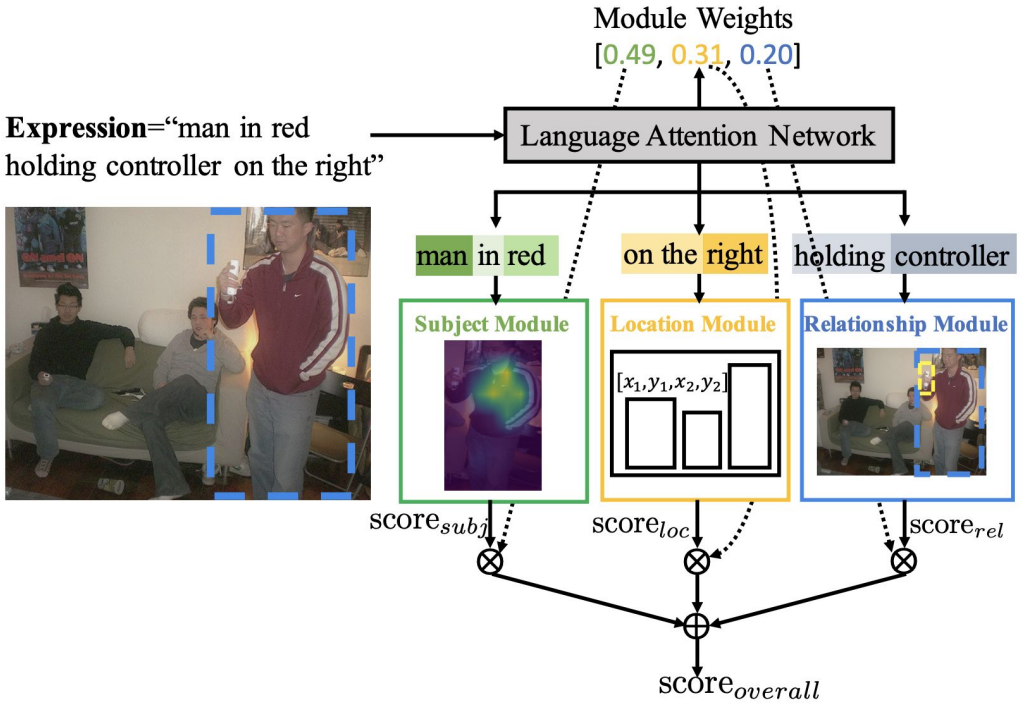
RefCOCOg:
1. an adult giraffe scratching its back with its horn
2. giraffe hugging another giraffe

Referring image segmentation: techniques

- MAttNet: Modular Attention Network for Referring Expression Comprehension [CVPR2018]
  - Motivation: most work treats 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 [CVPR2018]

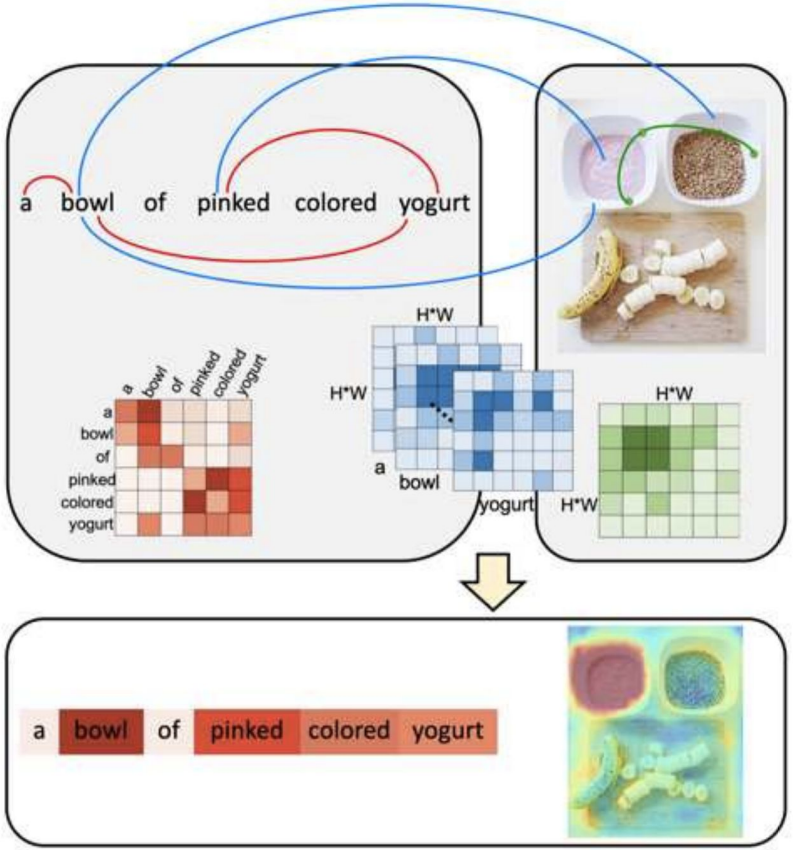


Referring image segmentation: techniques

- 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

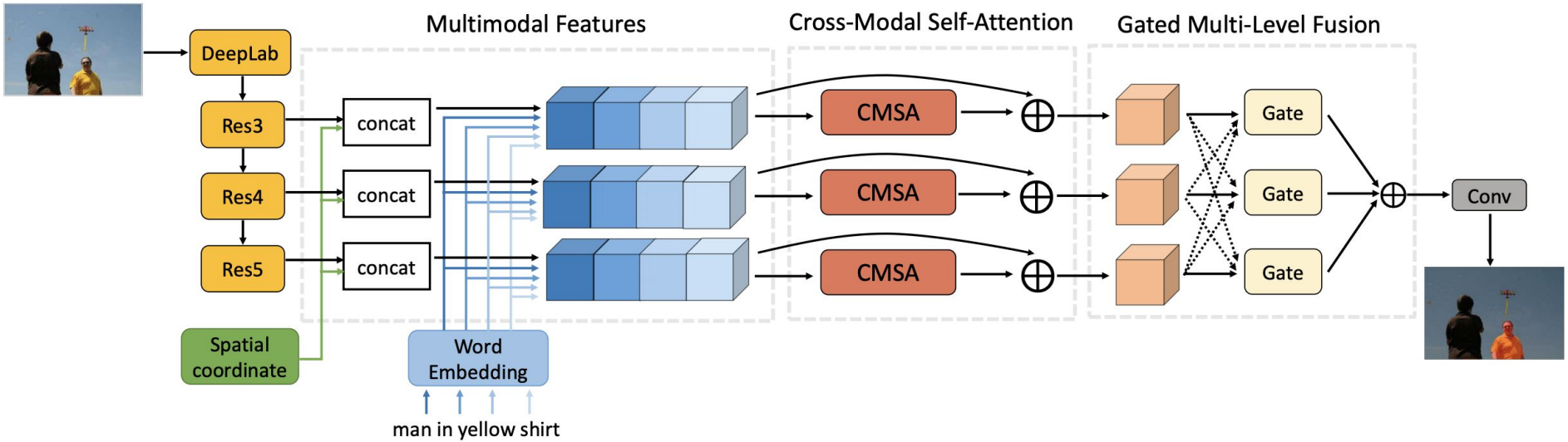
Referring image segmentation: techniques

- 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]

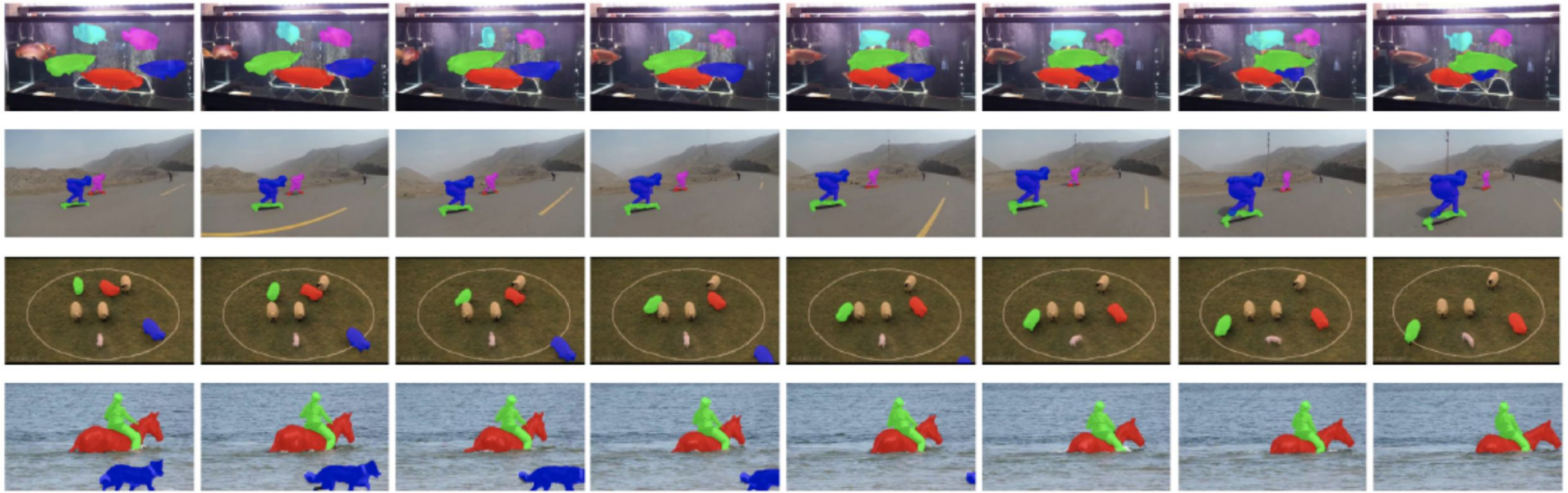


Outline

- 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]



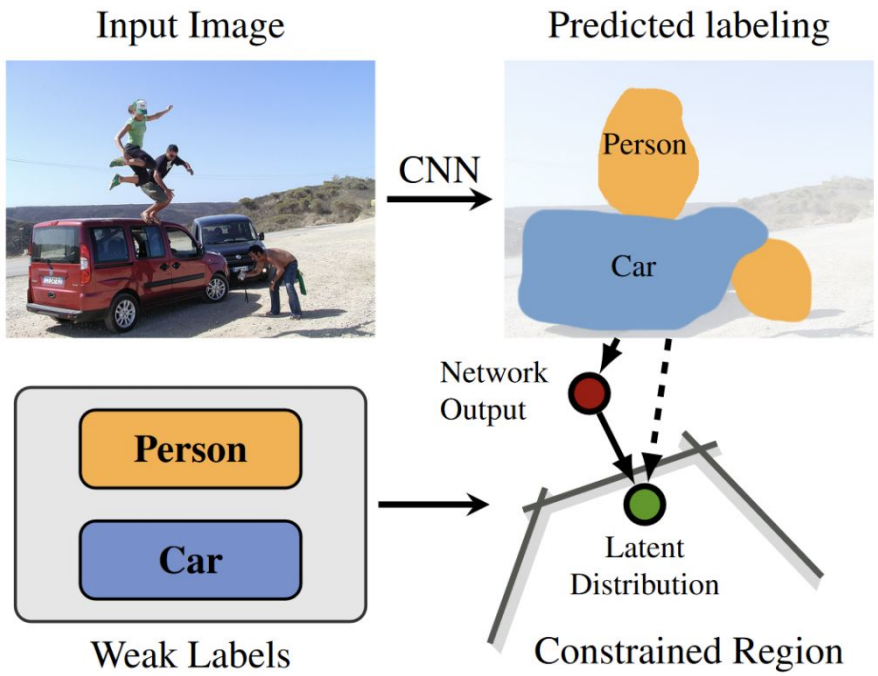
Current trends and future research

- Weakly supervised segmentation
  - How much does it cost to label this image? Money? Time?
    - Heavy labeling efforts
    - Pixel-wise labeling is expensive
    - Error-prone and hard to be precise
    - Categories can be too numerous…

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

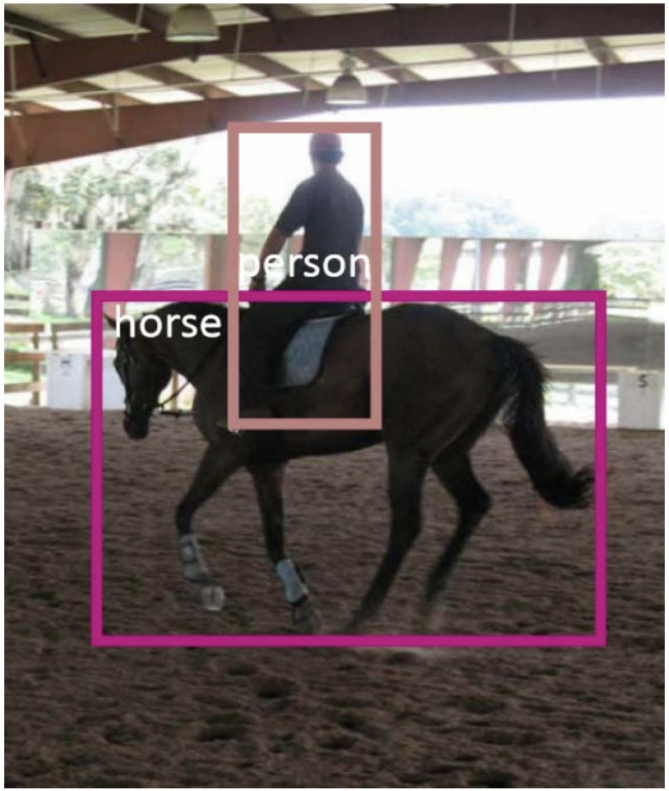


Current trends and future research

- 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

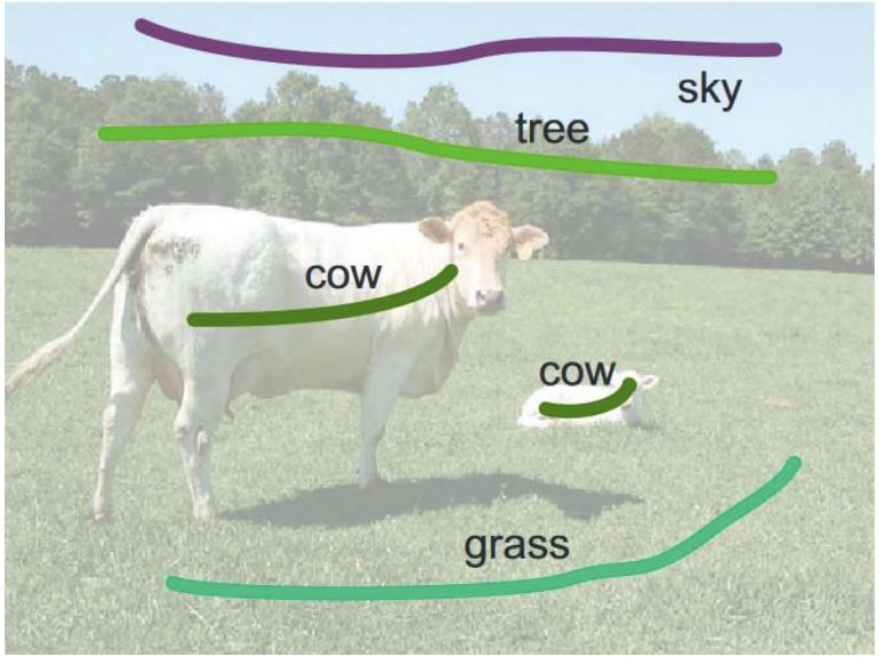
Current trends and future research

- 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



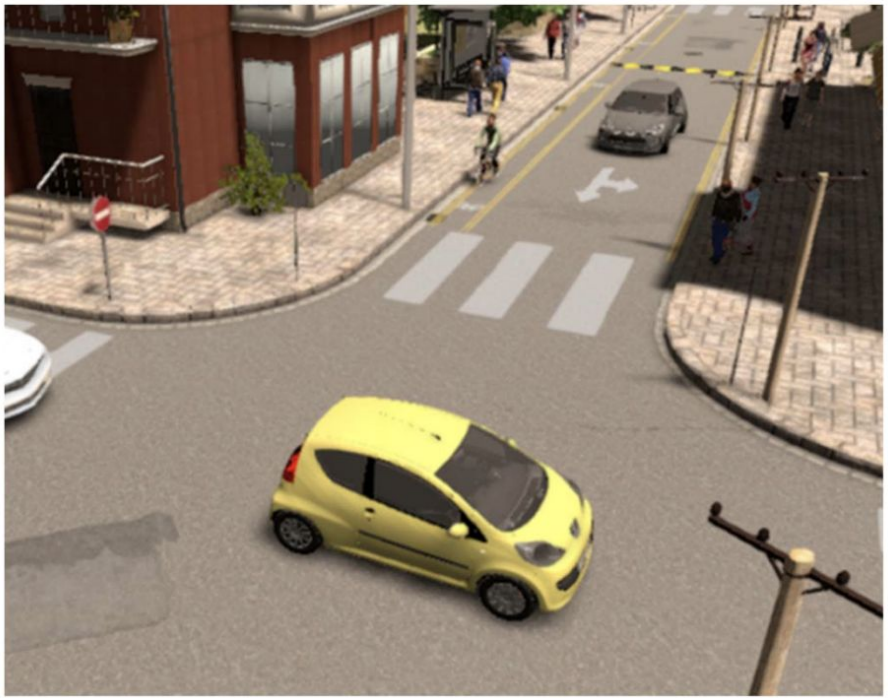
Current trends and future research

- 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

Semantic and Instance Segmentation

THANKS FOR YOUR ATTENTION

David Vazquez

ServiceNow

david.vazquez@servicenow.com

*Acknowledgements for slides credits to researchers

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