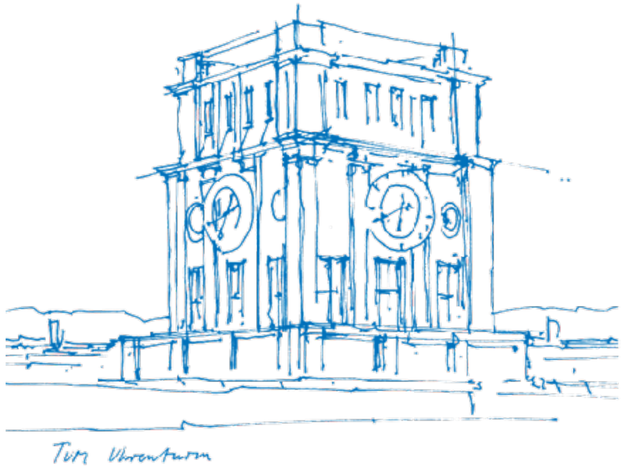
Computer Vision III:
Panoptic segmentation

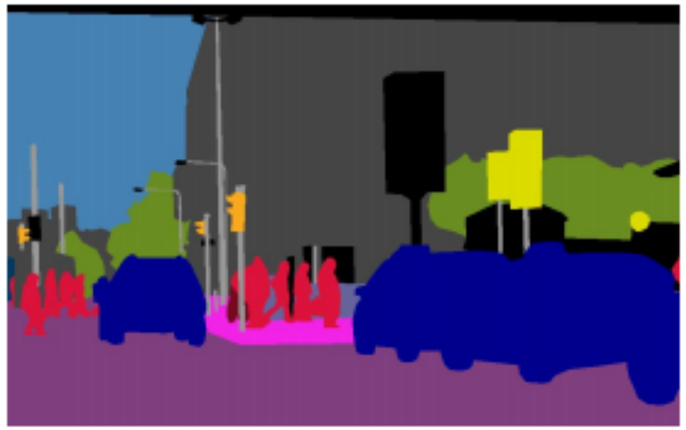
Nikita Araslanov
20.12.2022

Adapted from:
Prof. Laura Leal-Taixé
https://dvl.in.tum.de



Panoptic segmentation

Semantic segmentation

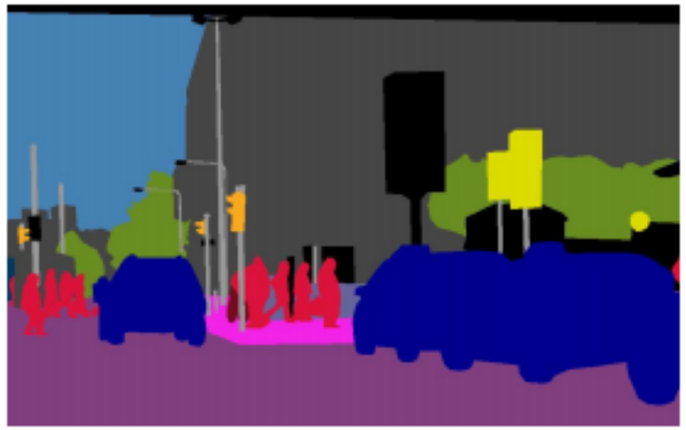


+

Instance segmentation

Panoptic segmentation

Semantic segmentation



(e.g. FCN, DeepLab)

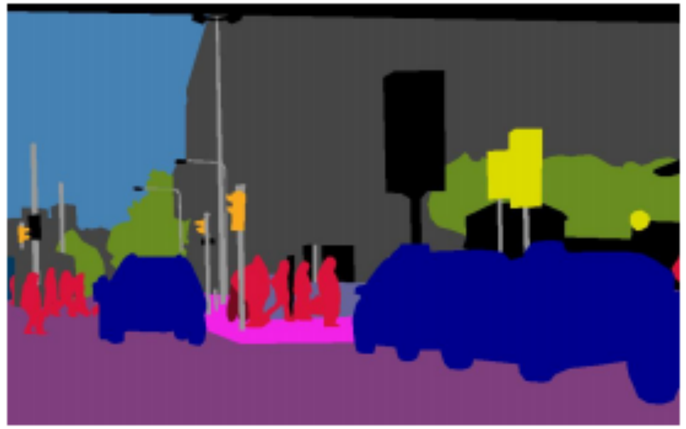
+

Instance segmentation

(e.g. Mask R-CNN)

Panoptic segmentation

Semantic segmentation



(e.g. FCN, DeepLab)

+

Instance segmentation

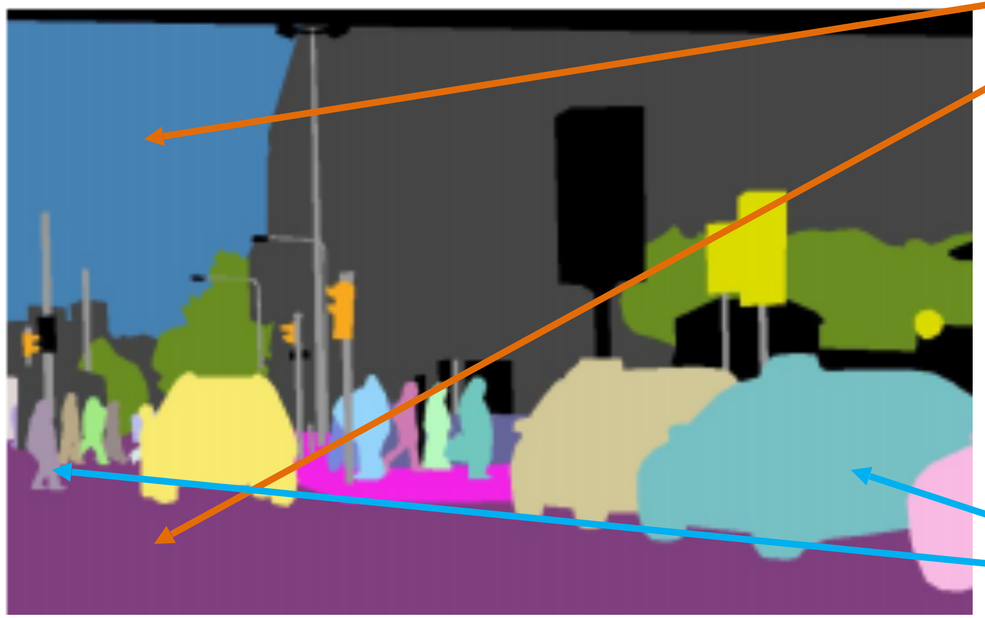
(e.g. Mask R-CNN)

=

Panoptic segmentation

(e.g. UPSNet)

Panoptic segmentation



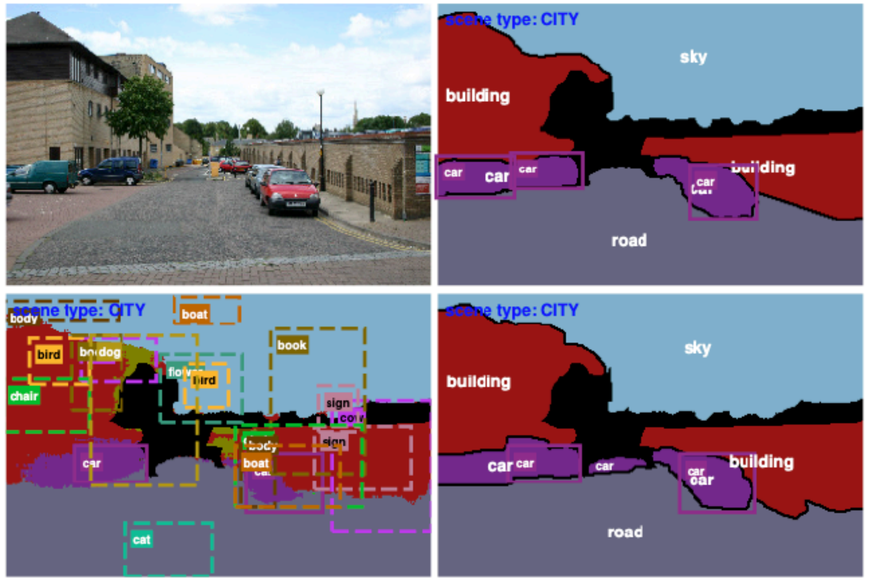
It gives labels to uncountable objects called "stuff" (sky, road, etc), similar to FCN-like networks.

It differentiates between pixels coming from different instances of the same class (countable objects) called "things" (cars, pedestrians, etc).

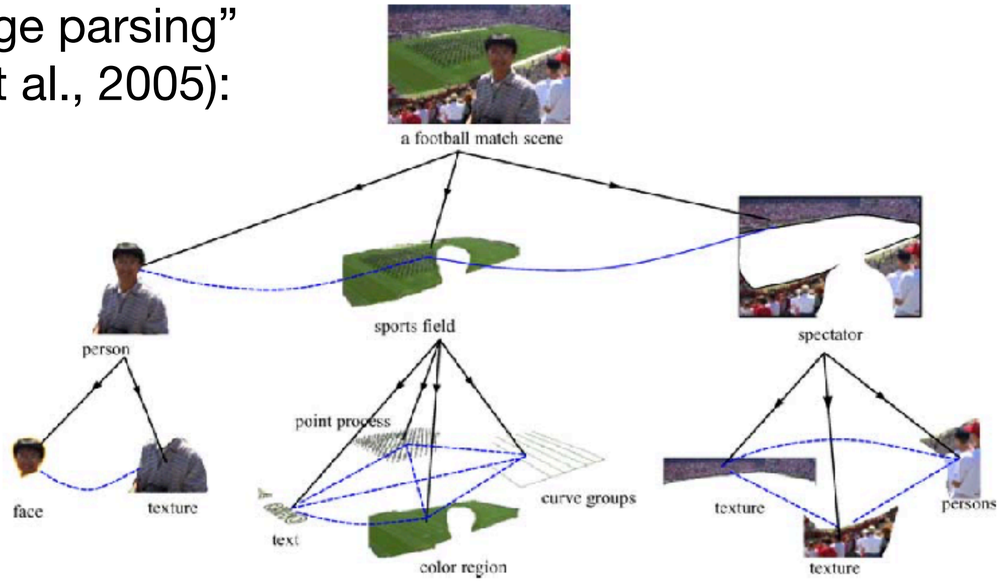
Back in the day

- The task is not new...

"Holistic scene understanding"
(Yao et al., 2012):



"Image parsing"
(Tu et al., 2005):



... but deep learning makes it feasible.

Panoptic segmentation

Challenges:

- Can we harmonise architectures for predicting “stuff” and “things”?
  - semantic and instance segmentation pipelines are yet very different.
- Can we improve computational efficiency via parameter sharing?

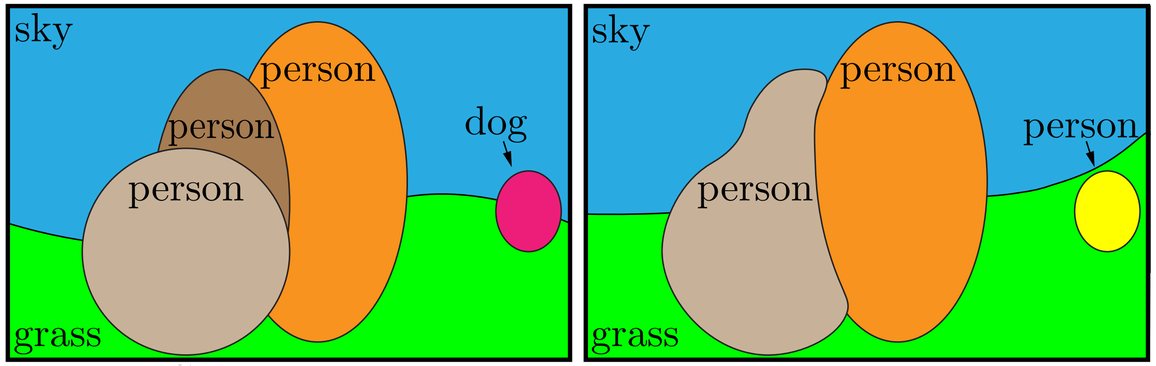
Two broad categories:

- Top-down: typically two-stage proposal-based.
- Bottom-up: learn suitable feature representation for grouping pixels.

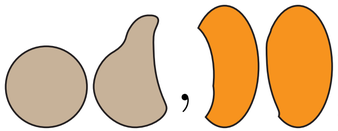
Evaluating panoptic segmentation

Panoptic quality (PQ)

- Example:



Ground truth Prediction

Person — TP: {  }; FN: {  }; FP: {  }

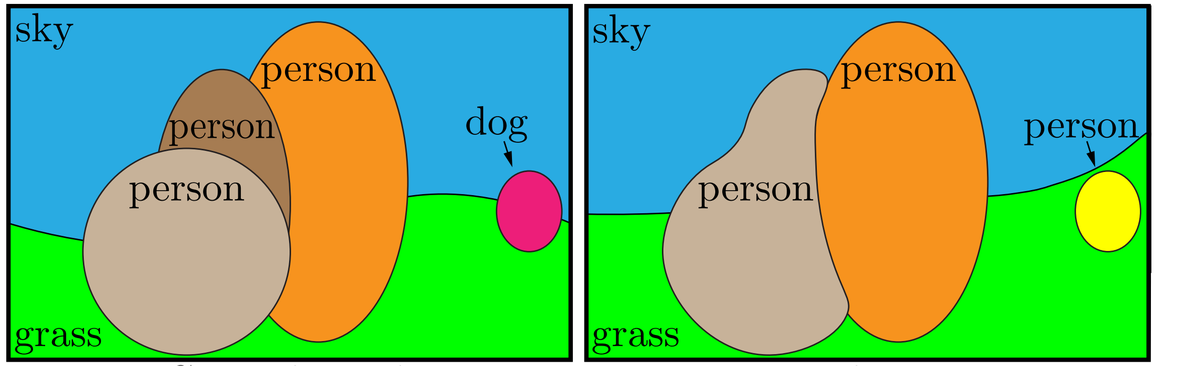
TP = True positive, FN = False negative, FP = false positive

- Wait, but don't we need to define an IoU threshold?

Kirillov et al., "Panoptic Segmentation". CVPR 2019.

Panoptic quality (PQ)

- To compute PQ we specify that a prediction and a ground truth match only if their IoU is greater than 0.5.
- This match, if found, is **unique**.
- Unique matching theorem:
  - A ground-truth segment has an IoU greater than 0.5 with at most **one** prediction.



proof hint

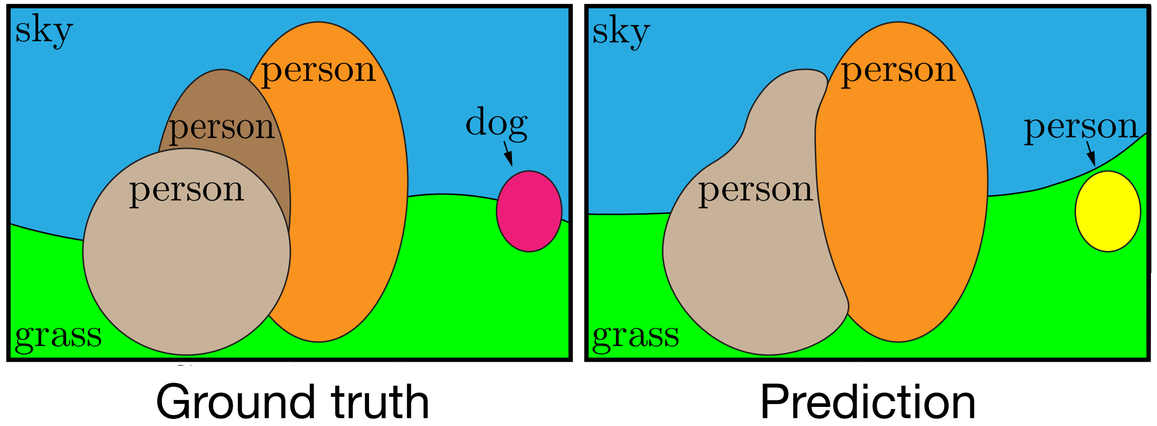
- With p_1 and p_2 being disjoint:

$$\mathrm{IoU}(p_1, g) + \mathrm{IoU}(p_2, g) \leq 1$$

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

1. Establish matches between the ground-truth and predictions;
2. Count TPs, FPs and FNs;
3. Compute PQ for each class:



Ground truth Prediction

Person — TP: { , }; FN: { }; FP: { }

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

...and then average.

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

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

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

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

- SQ = “Segmentation Quality”:
  - Average mask IoU for true positives;
  - Measures pixel-level accuracy of predicted masks.

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

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

RQ

- RQ = “Recognition Quality”:
  - Object-level accuracy.
  - Does it look familiar? (QUIZ)

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

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

RQ

- RQ = “Recognition Quality”:
  - Object-level accuracy.
  - Does it look familiar? This is F-score (F_1)
  - F-score is the harmonic mean of precision and recall.

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

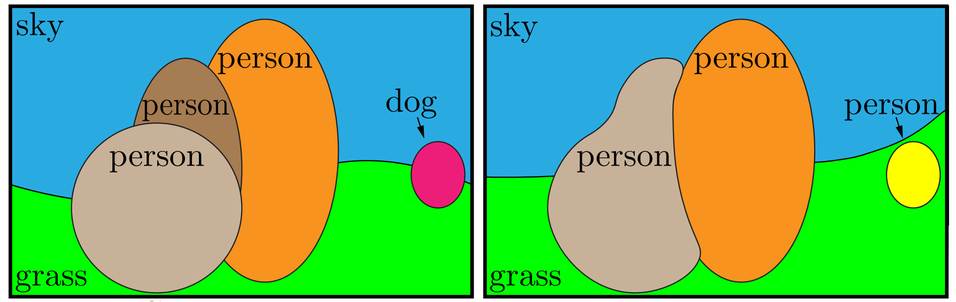
$$PQ = \frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\ \mathrm{TP}\ |} \frac{|\ \mathrm{TP}\ |}{|\ \mathrm{TP}\ | + \frac{1}{2}|\ \mathrm{FP}\ | + \frac{1}{2}|\ \mathrm{FN}\ |}$$

- Observation 1: $PQ, RQ, SQ \in [0, 1]$

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

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



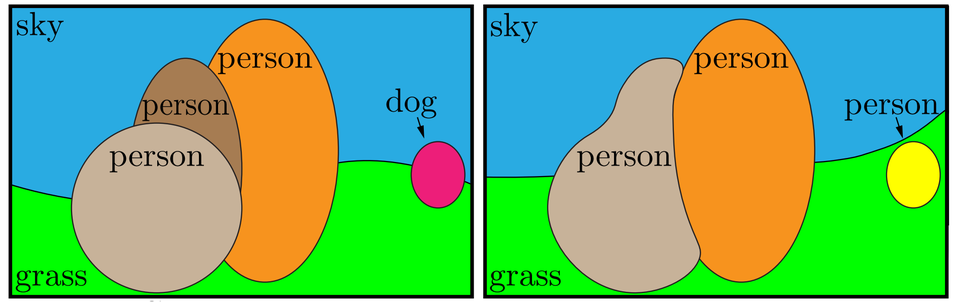
Ground truth Prediction

- Observation 2: What effect does missing one object have on PQ (e.g. “dog” above)?
  - Increment FN for that class (e.g. “dog”) AND FP for another class (e.g. “person”).

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

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



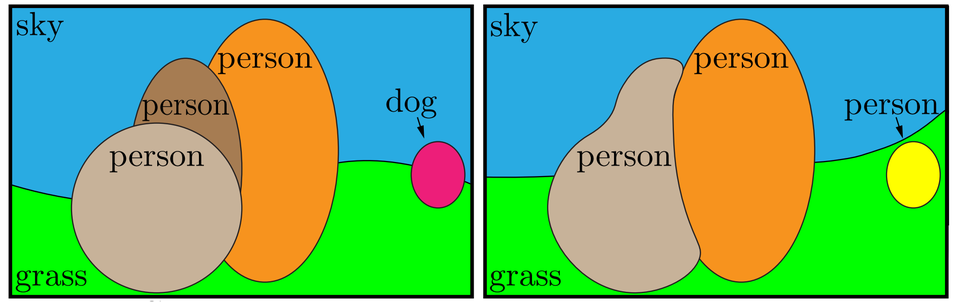
Ground truth Prediction

- Observation 2: What effect does missing one object have on PQ (e.g. “dog” above)?
  - Increment FN for that class (e.g. “dog”) AND FP for another class (e.g. “person”).
  - This reduces PQ for **two** classes.

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

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



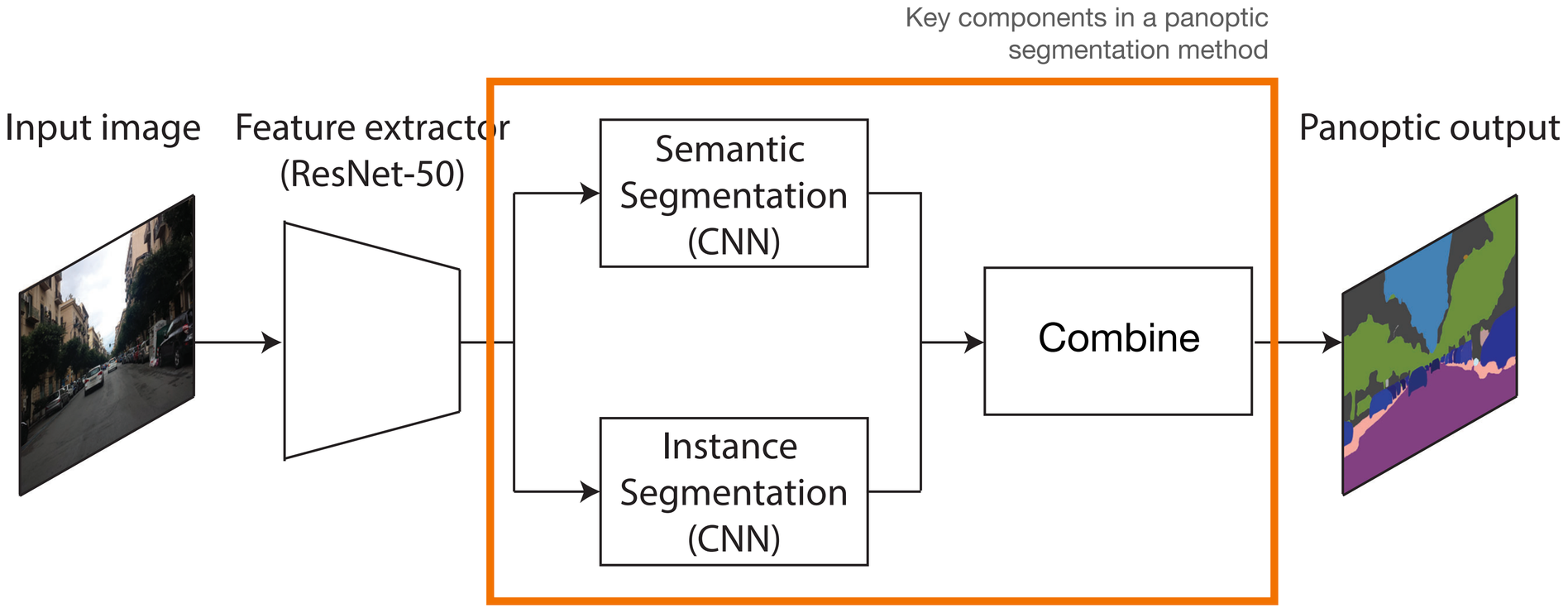
Ground truth Prediction

- Observation 2: What effect does missing one object have on PQ (e.g. “dog” above)?
  - Increment FN for that class (e.g. “dog”) AND FP for another class (e.g. “person”).
  - This reduces PQ for **two** classes.
  - Idea: Predict as “unknown” class instead. FP count will not affect another class.

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Overview

- Typical architecture:



Adapted from [de Geus et al., 2018].

Overview

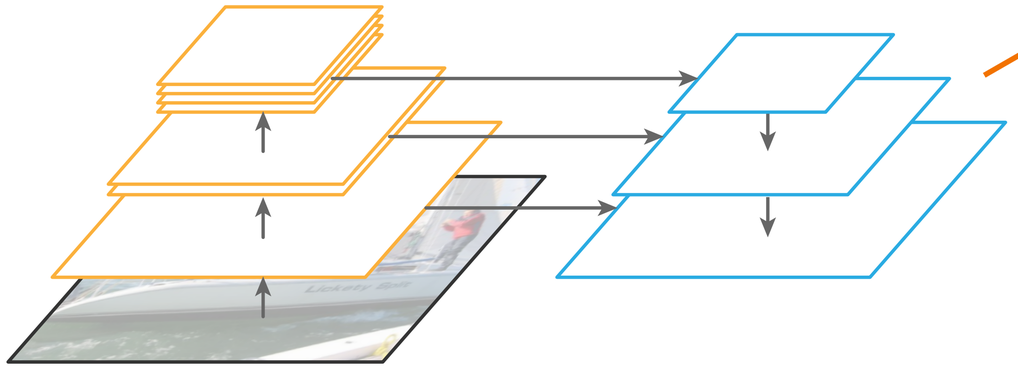
- Kirillov et al., “Panoptic Feature Pyramid Networks”, CVPR 2019.
- Xiong et al., “UPSNet: A Unified Panoptic Segmentation Network”, CVPR 2019.
- Cheng et al., “Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation”, CVPR 2020.
- Li et al., “Fully Convolutional Networks for Panoptic Segmentation”, CVPR 2021.

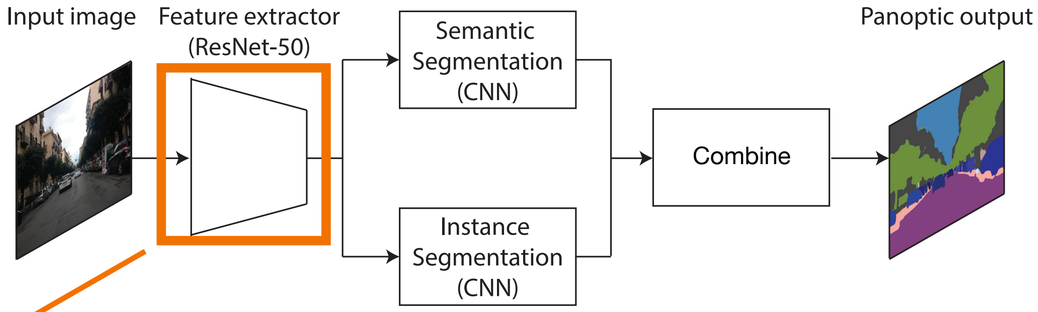
Overview

- **Kirillov et al., “Panoptic Feature Pyramid Networks”, CVPR 2019.**
- Xiong et al., “UPSNet: A Unified Panoptic Segmentation Network”, CVPR 2019.
- Cheng et al., “Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation”, CVPR 2020.
- Li et al., “Fully Convolutional Networks for Panoptic Segmentation”, CVPR 2021.

Panoptic FPN

- Feature pyramid backbone:

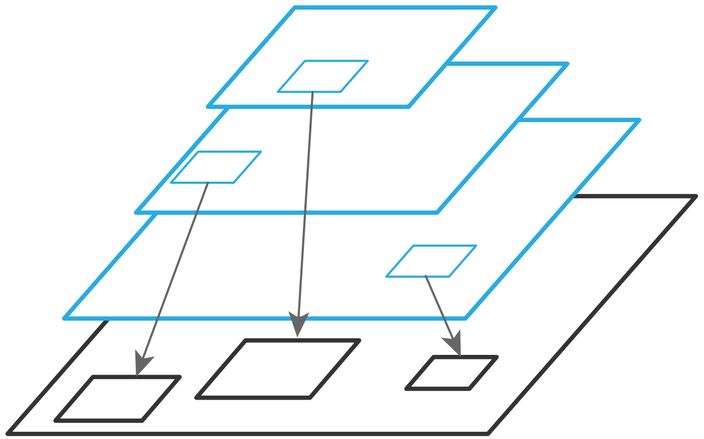


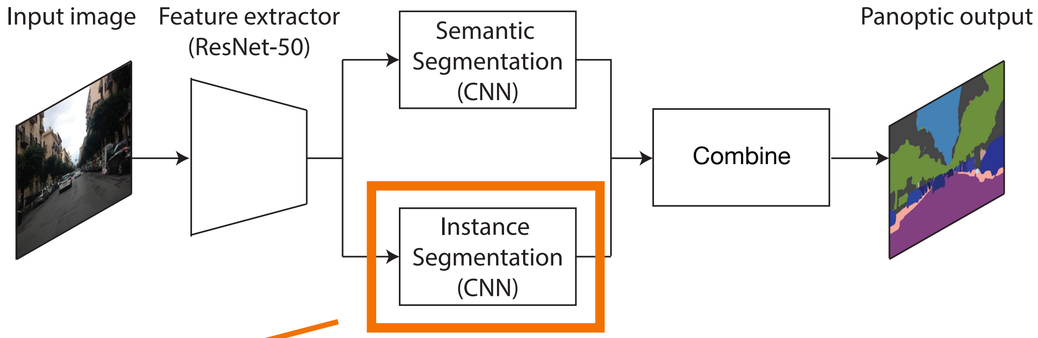


Kirillov et al., “Panoptic Feature Pyramid Networks”. CVPR 2019

Panoptic FPN

- Mask R-CNN for instance segmentation

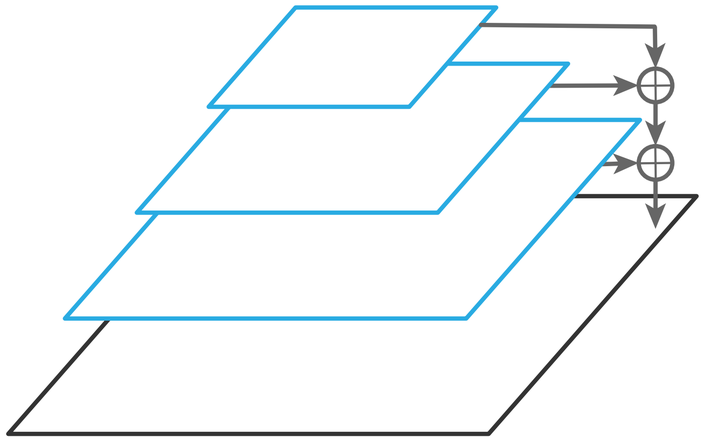


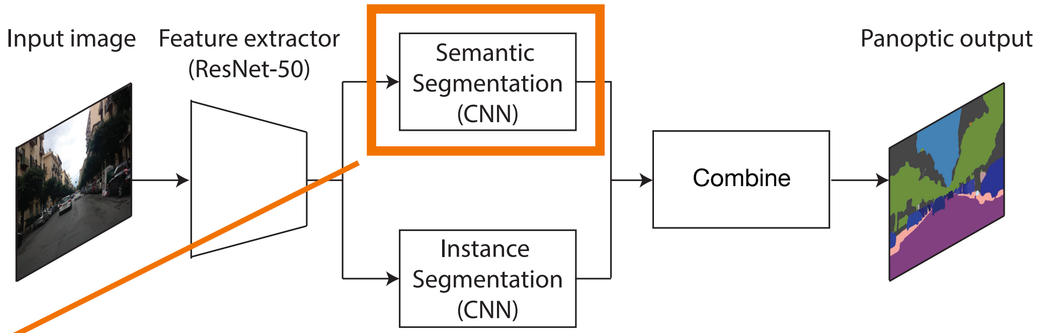


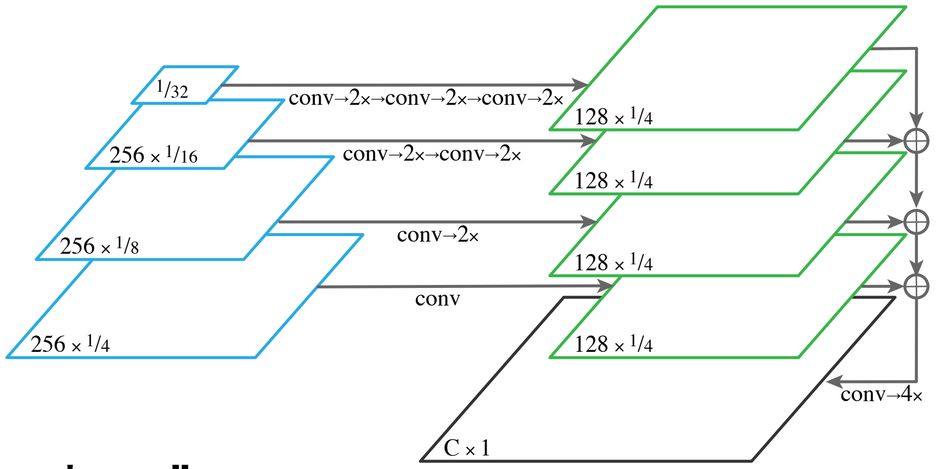
Kirillov et al., “Panoptic Feature Pyramid Networks”. CVPR 2019

Panoptic FPN

- Semantic segmentation decoder:





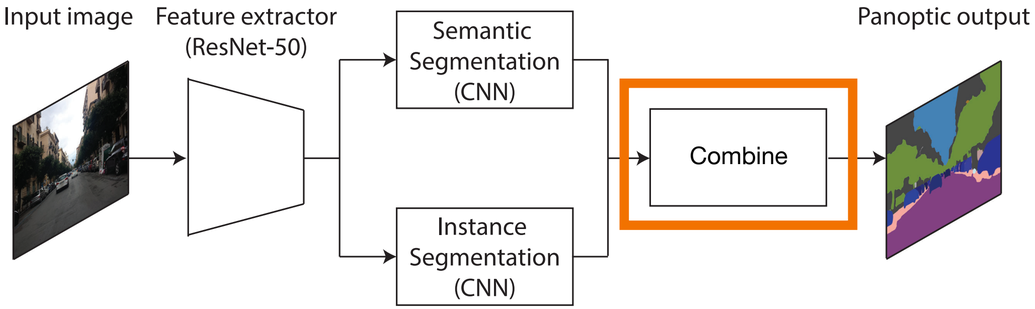


- Replace things classes with 1 class “other”

Kirillov et al., “Panoptic Feature Pyramid Networks”. CVPR 2019

Panoptic FPN

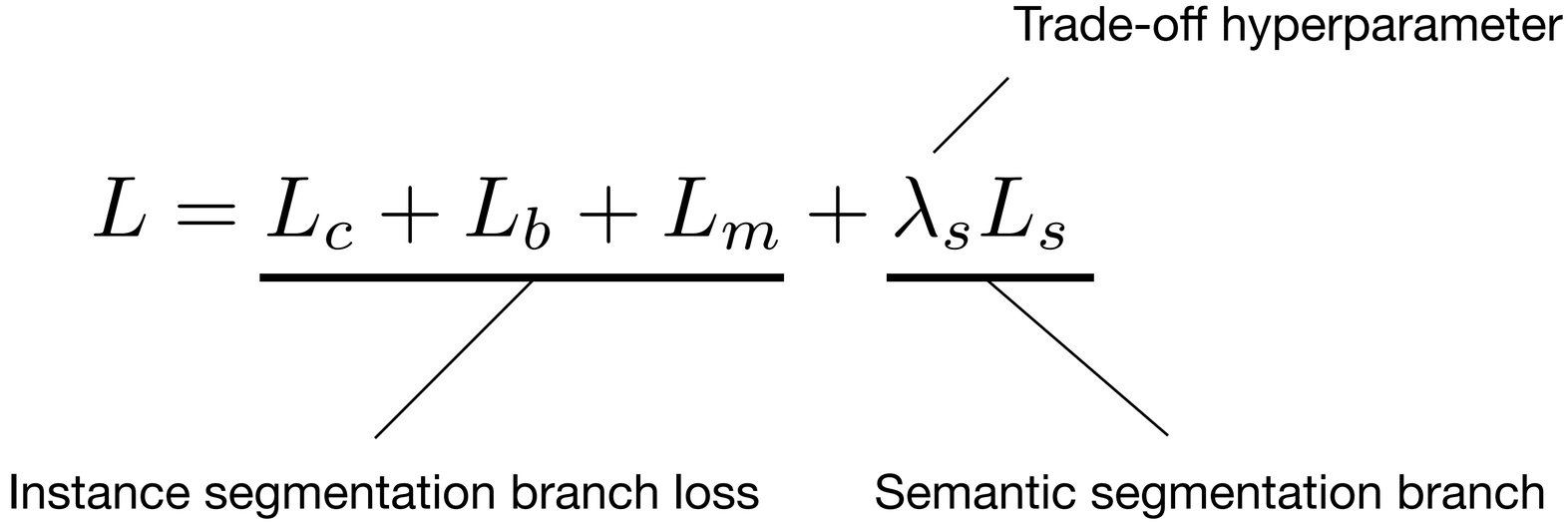
- Merge things and stuff:
  - NMS on instances.
  - Resolve stuff-things conflicts in favour of things. (Why?)
  - Remove any stuff regions labelled “other” or with a small area.



Kirillov et al., “Panoptic Feature Pyramid Networks”. CVPR 2019

Panoptic FPN

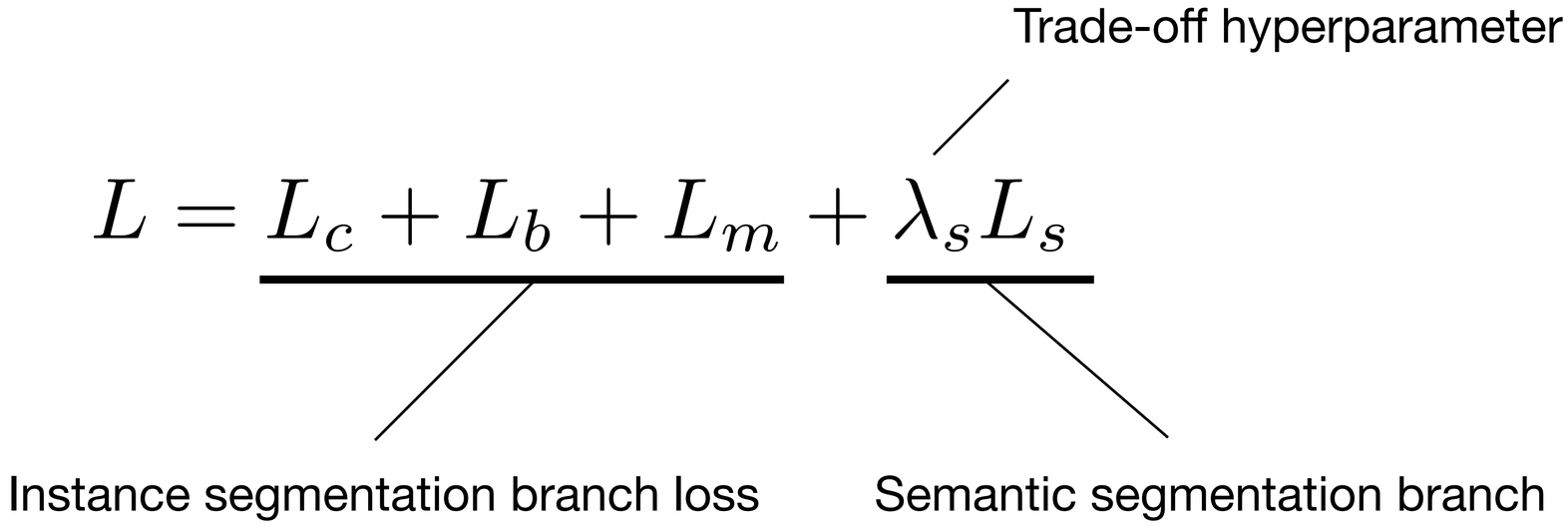
- Loss function:



Kirillov et al., “Panoptic Feature Pyramid Networks”. CVPR 2019

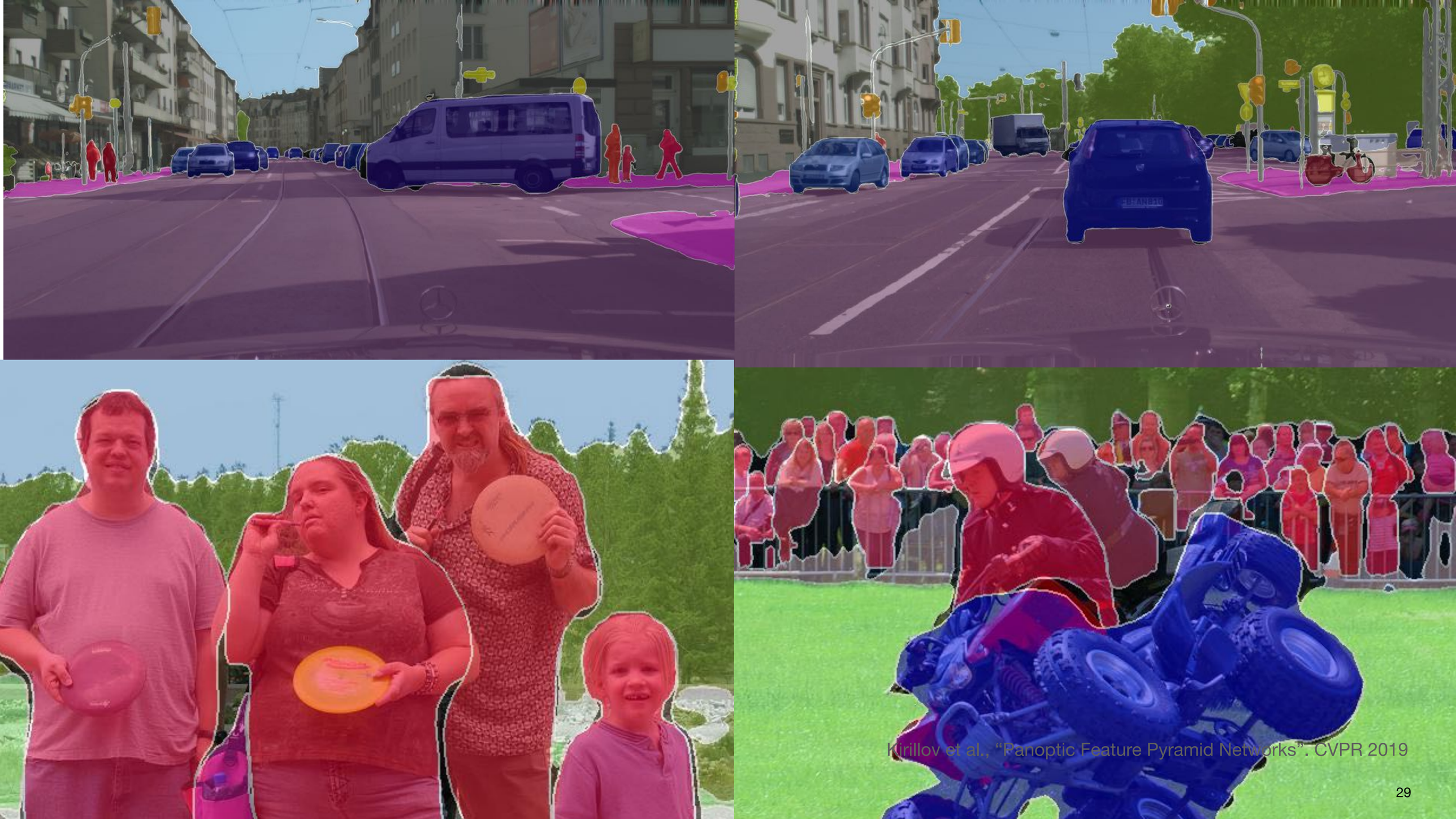
Panoptic FPN

- Loss function:



- Remark: Training with multiple loss terms (“multi-task learning”) can be challenging, as different loss terms may “compete” for desired feature representation.

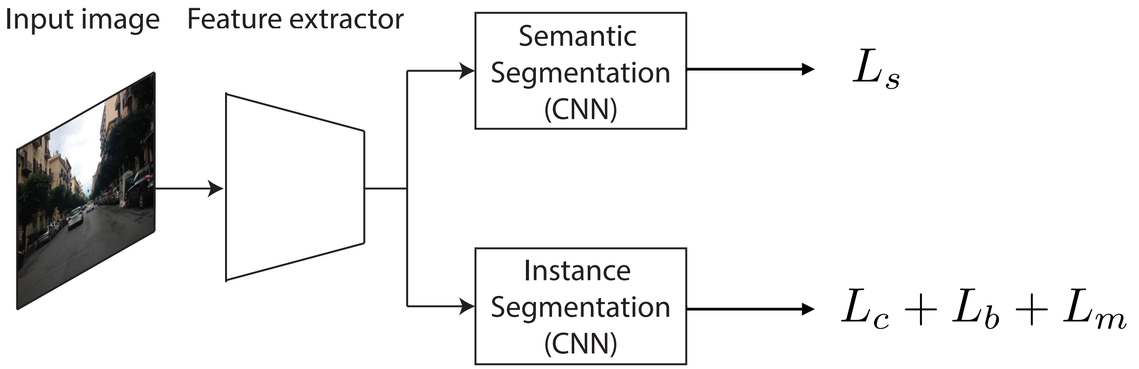
Kirillov et al., “Panoptic Feature Pyramid Networks”. CVPR 2019



Kirillov et al., “Panoptic Feature Pyramid Networks”. CVPR 2019

Panoptic FPN: Summary

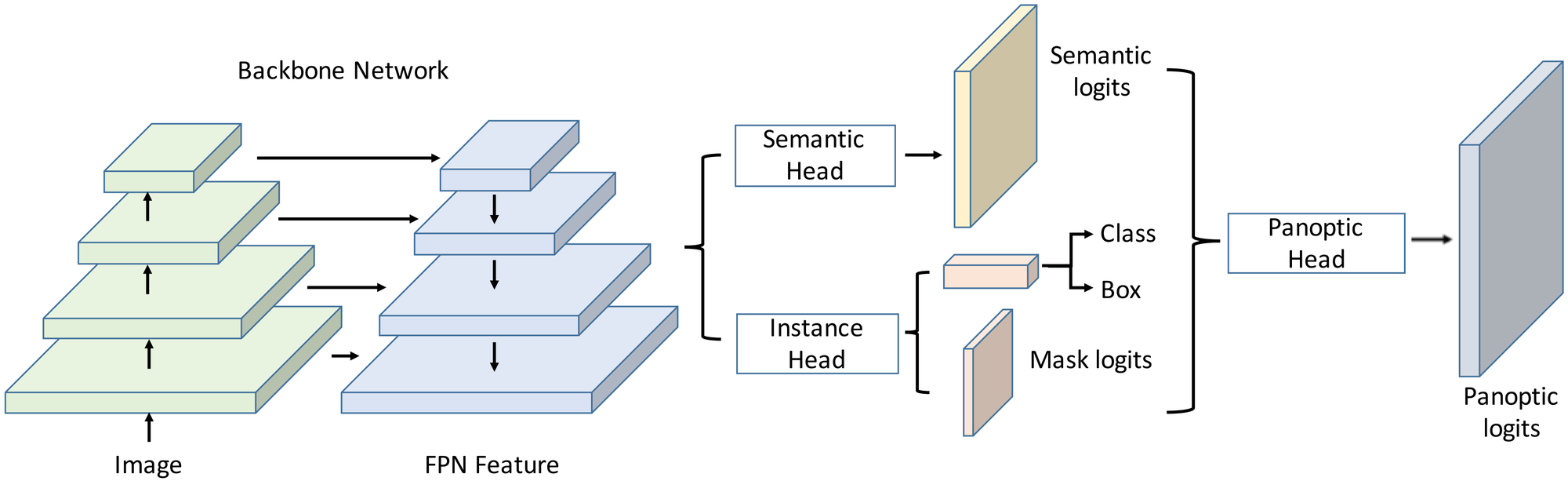
- Simple heuristics for merging things and stuff
- The instance and semantic segmentation branches are treated independently
  - i.e. semantic segmentation branch receives no gradient from instance supervision and vice-versa.



Overview

- Kirillov et al., “Panoptic Feature Pyramid Networks”, CVPR 2019.
- **Xiong et al., “UPSNet: A Unified Panoptic Segmentation Network”, CVPR 2019.**
- Cheng et al., “Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation”, CVPR 2020.
- Li et al., “Fully Convolutional Networks for Panoptic Segmentation”, CVPR 2021.

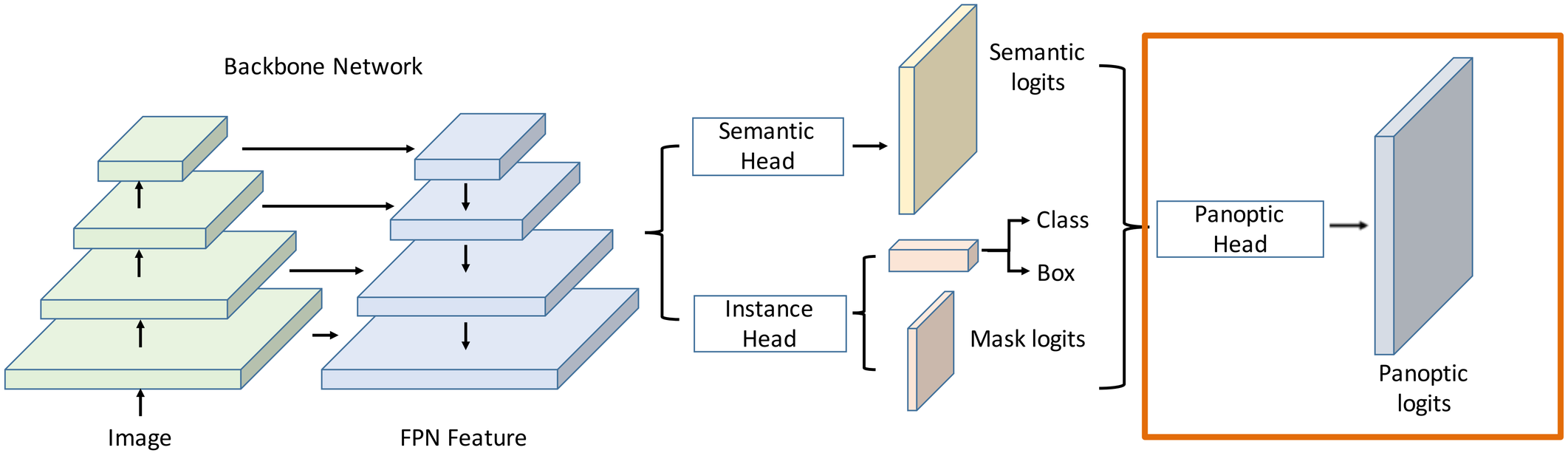
UPSNet



- What's different?

Xiong et al., "UPSNet: A Unified Panoptic Segmentation Network" (2019).

UPSNet

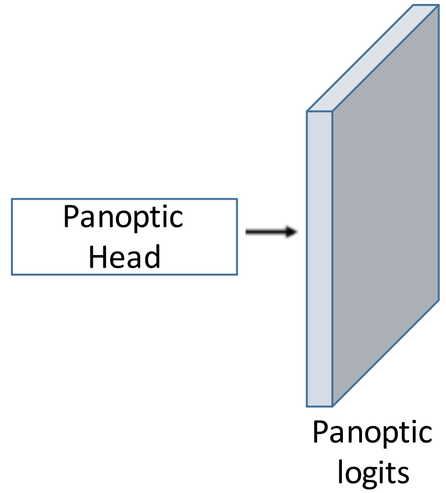


- What's different?

Xiong et al., "UPSNet: A Unified Panoptic Segmentation Network" (2019).

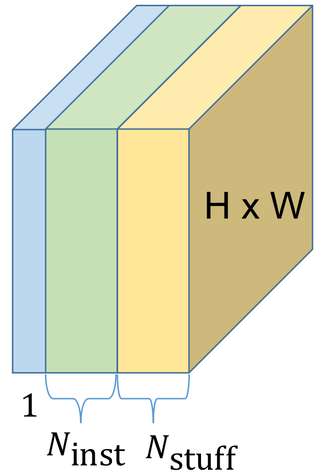
The panoptic head

- Main idea: create a single logit tensor for stuff and things
  - number of stuff categories: N_{stuff}
  - number of instances (variable per image): N_{inst}
  - add 1 “unknown” category
  - panoptic tensor: $(N_{\text{stuff}} + N_{\text{inst}} + 1) \times H \times W$



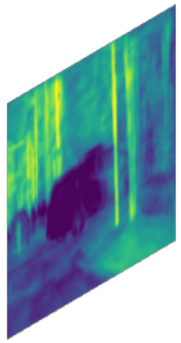
- Semantic segmentation head predicts both things and stuff classes.

The panoptic head



Panoptic logits

Stuff logits coming from the semantic head (e.g., sky)



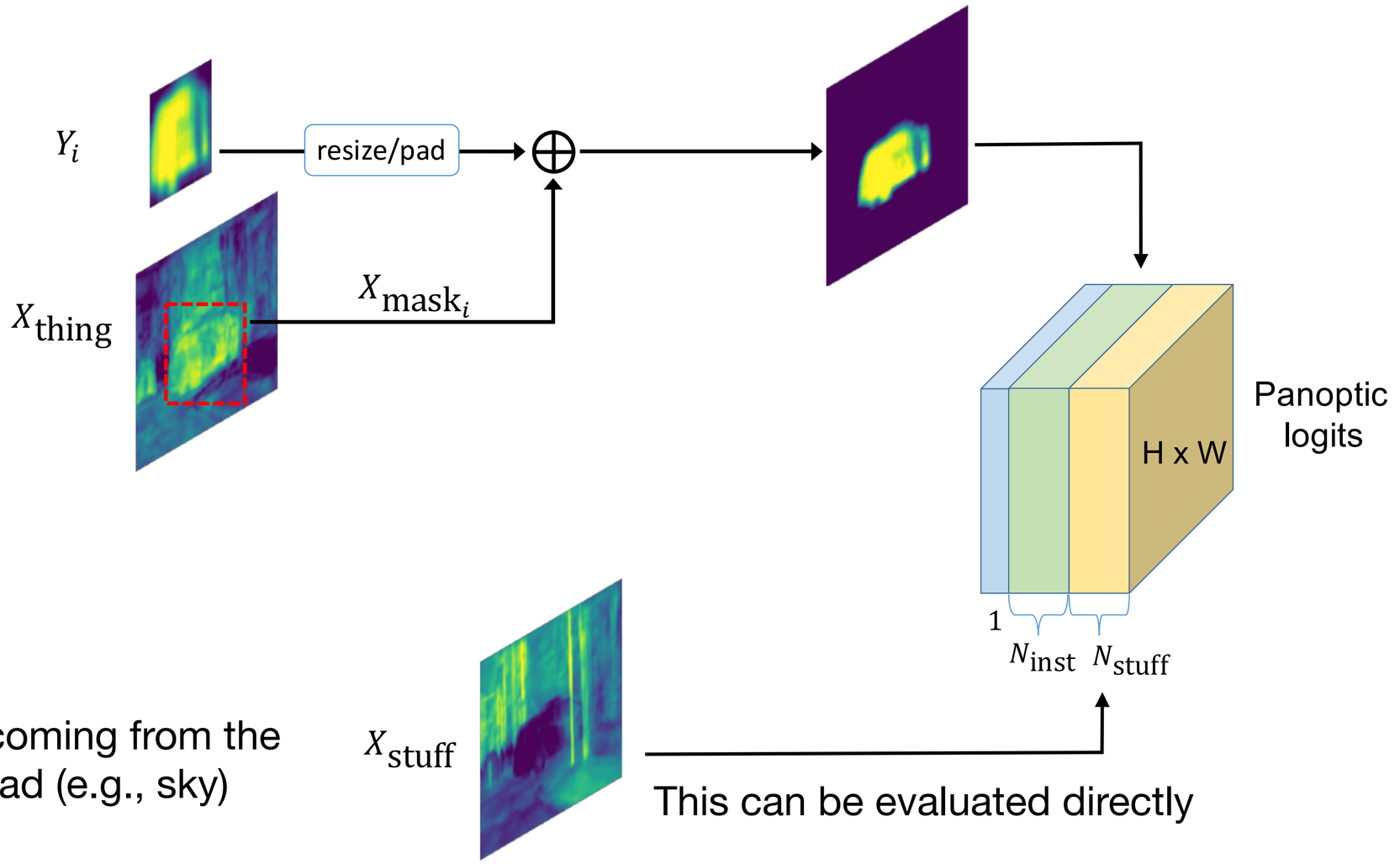
This can be evaluated directly

Xiong et al., “UPSNet: A Unified Panoptic Segmentation Network” (2019).

The panoptic head

Mask logits from the instance head

Object logits coming from the semantic head (e.g., car)



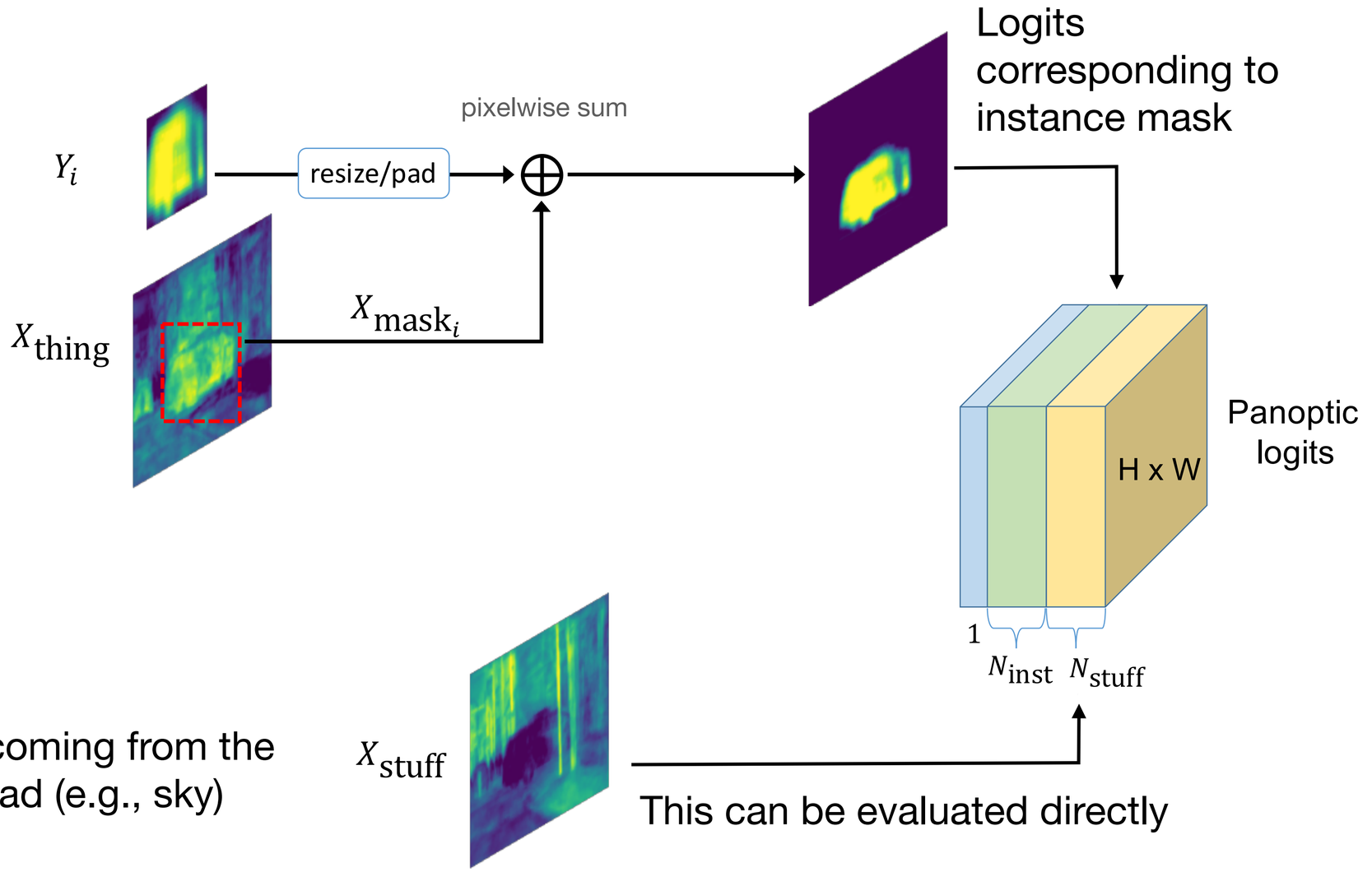
Stuff logits coming from the semantic head (e.g., sky)

Xiong et al., “UPSNet: A Unified Panoptic Segmentation Network” (2019).

The panoptic head

Mask logits from the instance head

Object logits coming from the semantic head (e.g., car)



Stuff logits coming from the semantic head (e.g., sky)

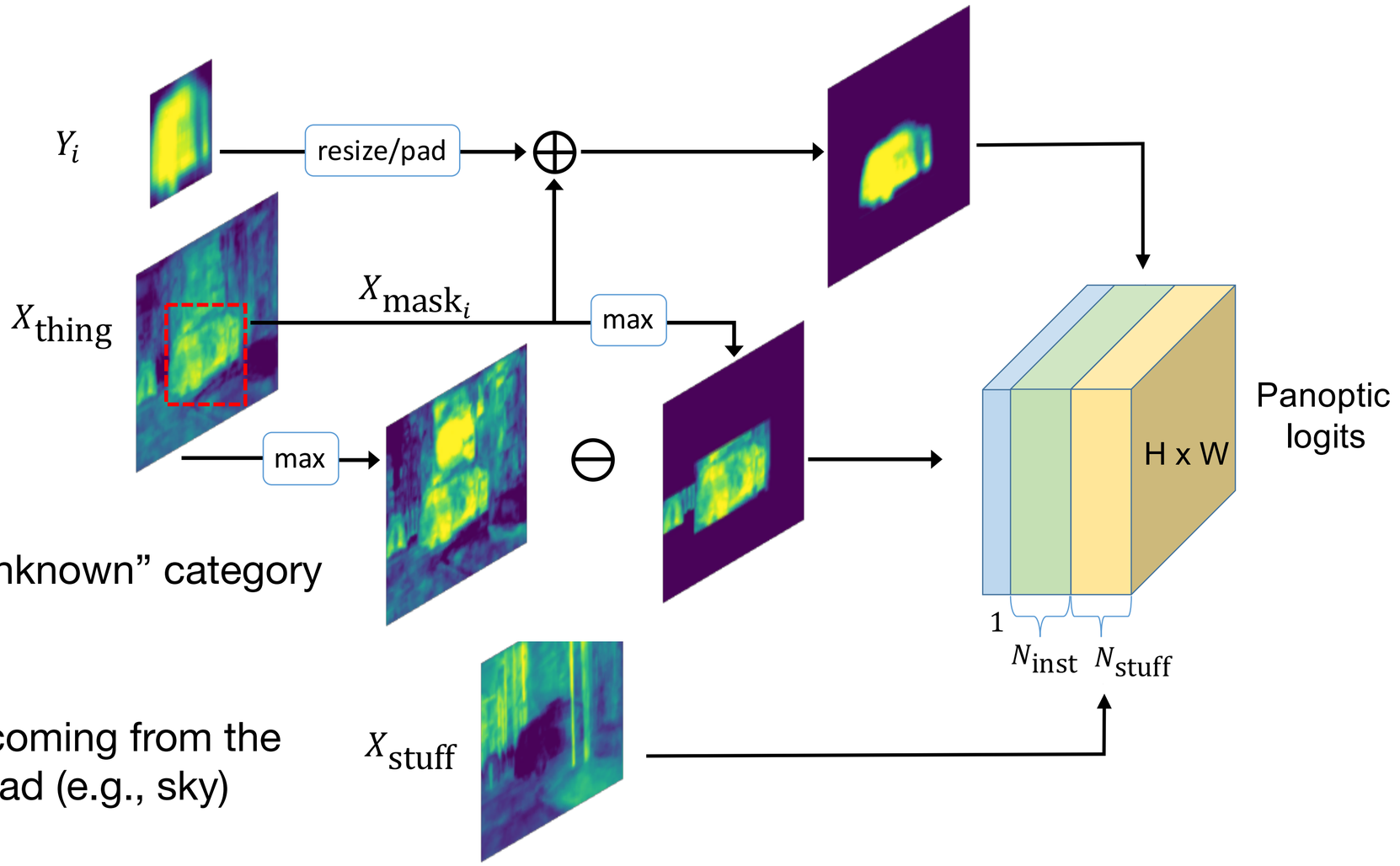
This can be evaluated directly

Xiong et al., “UPSNet: A Unified Panoptic Segmentation Network” (2019).

The panoptic head

Mask logits from the instance head

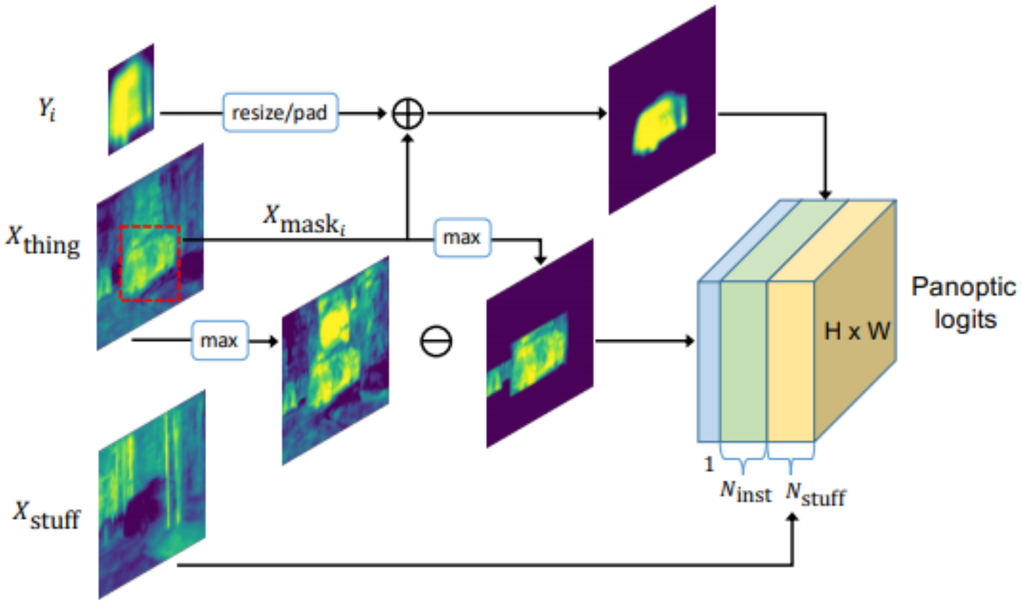
Object logits coming from the semantic head (e.g., car)



Stuff logits coming from the semantic head (e.g., sky)

Xiong et al., "UPSNet: A Unified Panoptic Segmentation Network" (2019).

The panoptic head

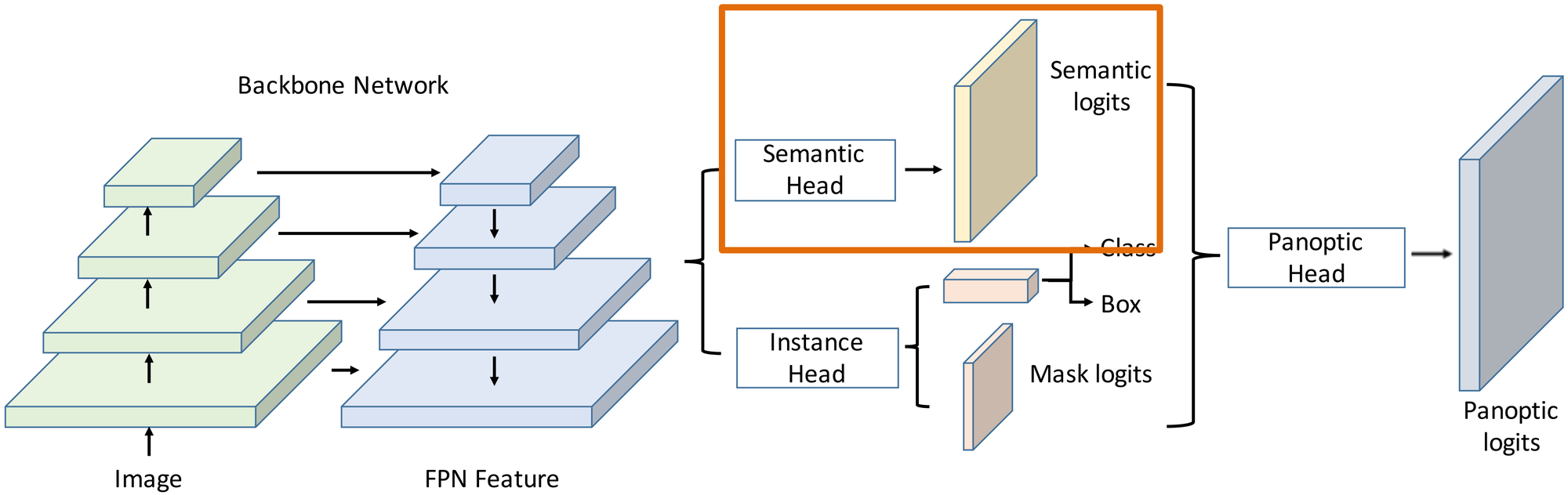


Advantage of panoptic logits:

- the channel argmax tells us if the pixel belongs to stuff, “unknown” or things;
- if the latter, which ID.

Xiong et al., “UPSNet: A Unified Panoptic Segmentation Network”. CVPR 2019

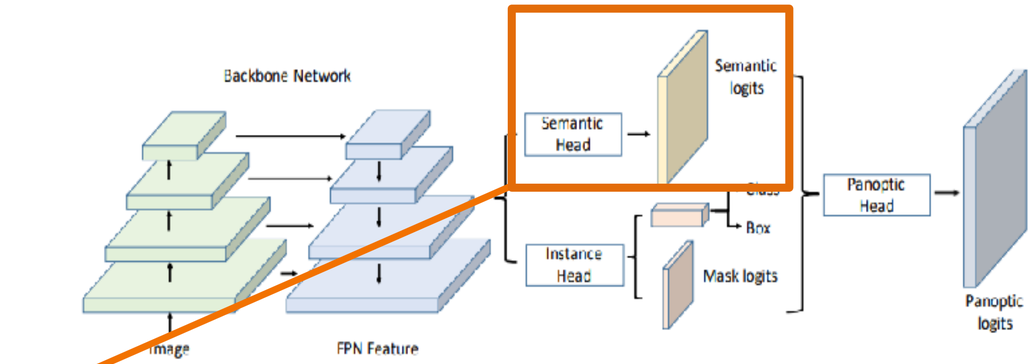
UPSNet

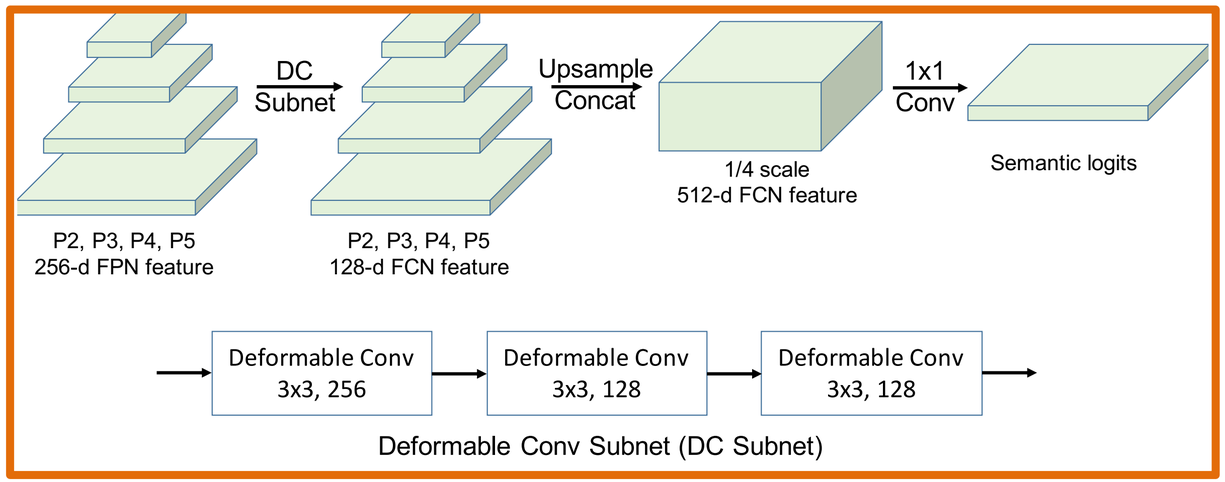


- What's different?

Xiong et al., "UPSNet: A Unified Panoptic Segmentation Network" (2019).

The semantic head

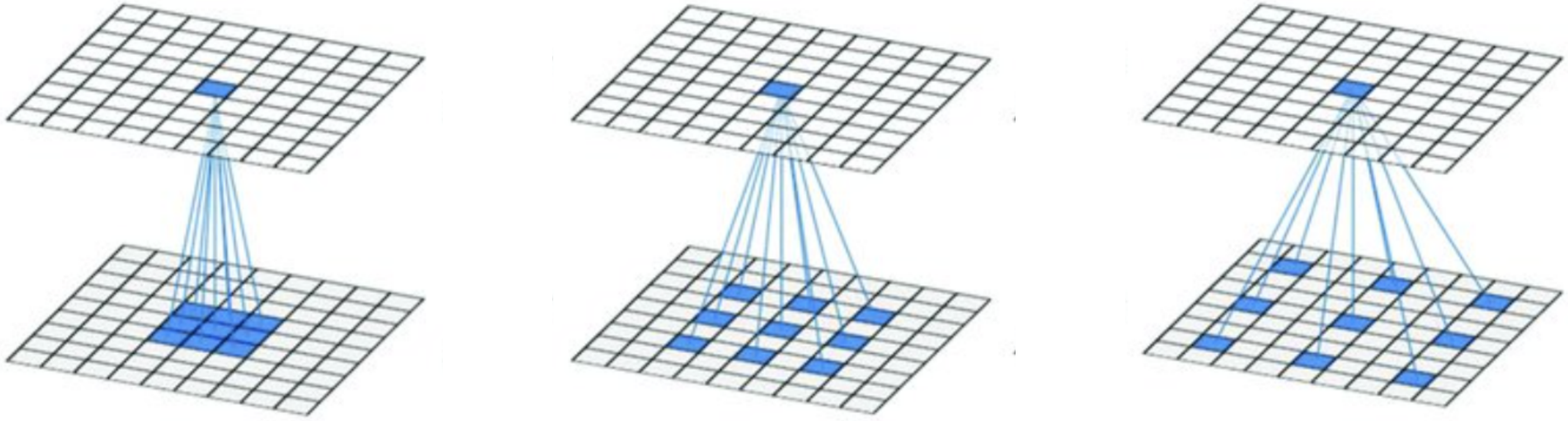




- New: deformable convolutions.

Recall: Dilated convolutions 2D

- Consider convolution as a special case with “dilation” = 1;
- For dilation N, the kernel “skips” N-1 pixels in-between:



dilation = 1 dilation = 2 dilation = 3

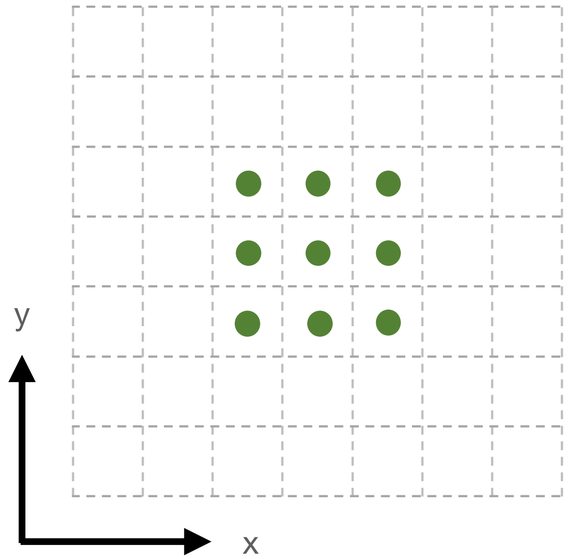
Generalising dilation

- We can represent dilated convolution as parameterisation of a non-dilated convolutional kernel:

$$k'(x, y) := k(x + g_x(x, y), y + g_y(x, y))$$

- Example:

3x3
dilation 1



(g_x, g_y)

0,0	0,0	0,0
0,0	0,0	0,0
0,0	0,0	0,0

Dai et al., “Deformable Convolutional Networks” ICCV 2017.

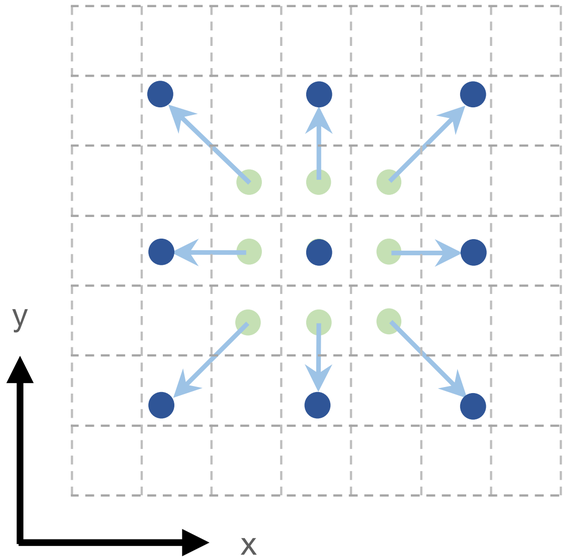
Generalising dilation

- We can represent dilated convolution as parameterisation of a non-dilated convolutional kernel:

$$k'(x, y) := k(x + g_x(x, y), y + g_y(x, y))$$

- Example:

3x3
dilation 2



(g_x, g_y)

-1,1	0,1	1,1
-1,0	0,0	1,0
-1,-1	0,-1	1,-1

Dai et al., “Deformable Convolutional Networks” ICCV 2017.

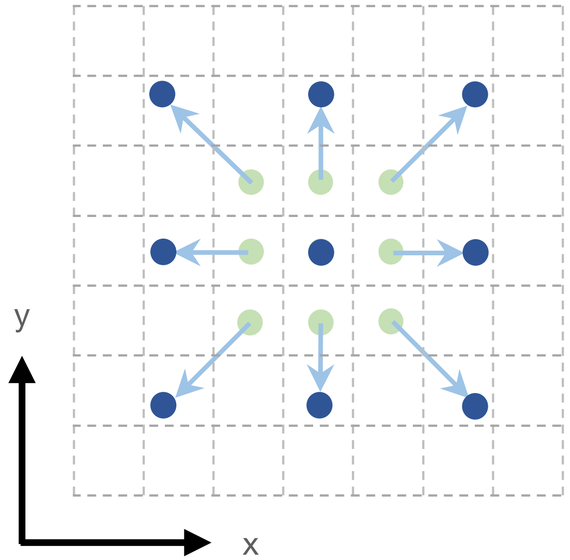
Generalising dilation

- We can represent dilated convolution as parameterisation of a non-dilated convolutional kernel:

$$k'(x, y) := k(x + g_x(x, y), y + g_y(x, y))$$

- Example:

3x3
dilation 2



(g_x, g_y)

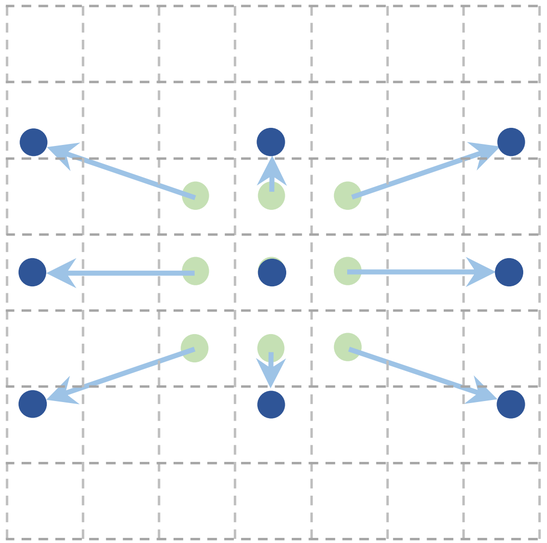
-1,1	0,1	1,1
-1,0	0,0	1,0
-1,-1	0,-1	1,-1

Can be fractional
(use bilinear interpolation)

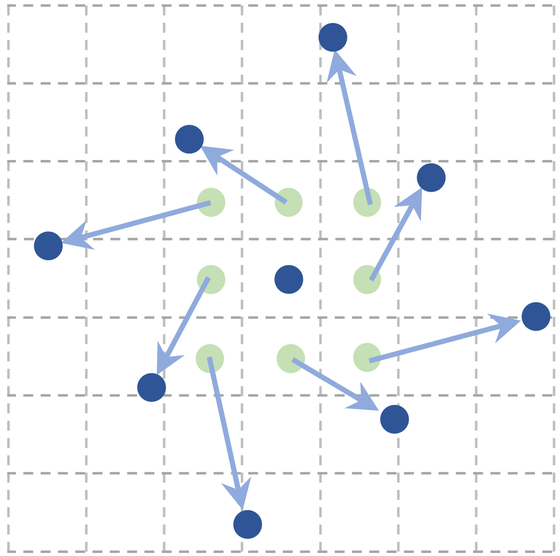
Dai et al., “Deformable Convolutional Networks” ICCV 2017.

Generalising dilation

- More examples:



Horizontal scaling

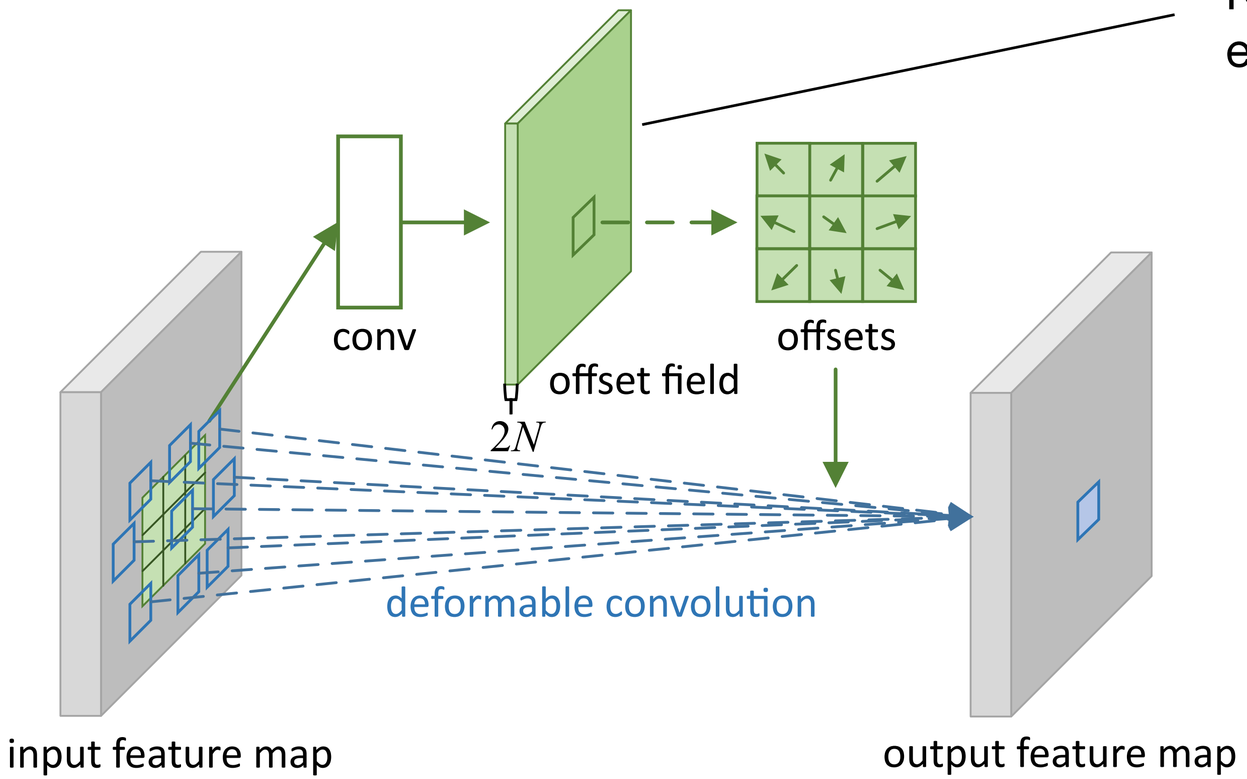


Rotation

Dai et al., “Deformable Convolutional Networks” ICCV 2017.

Deformable convolutions

- Learnable offsets:

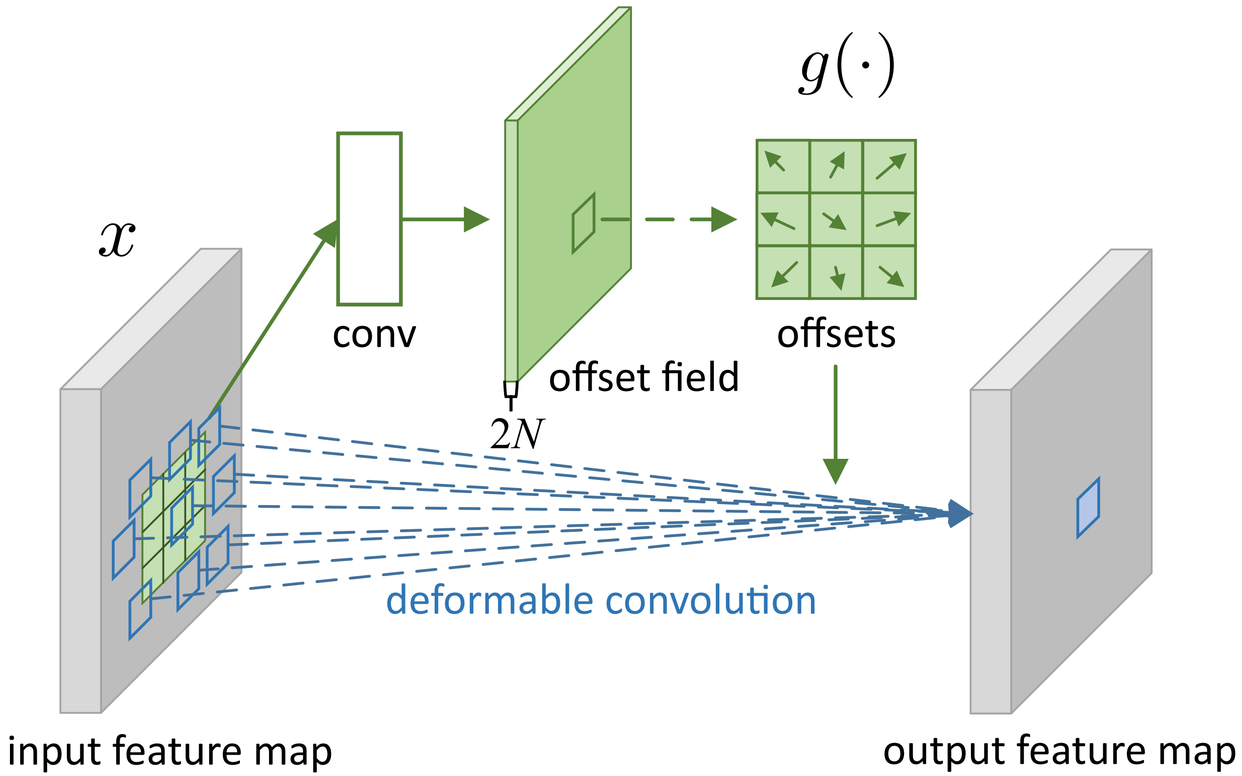


N is the number of offsets (kernel size); each offset is two scalars (x, y)

Dai et al., “Deformable Convolutional Networks” ICCV 2017.

Deformable convolutions

- Learnable offsets:



Regular convolution:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n)$$

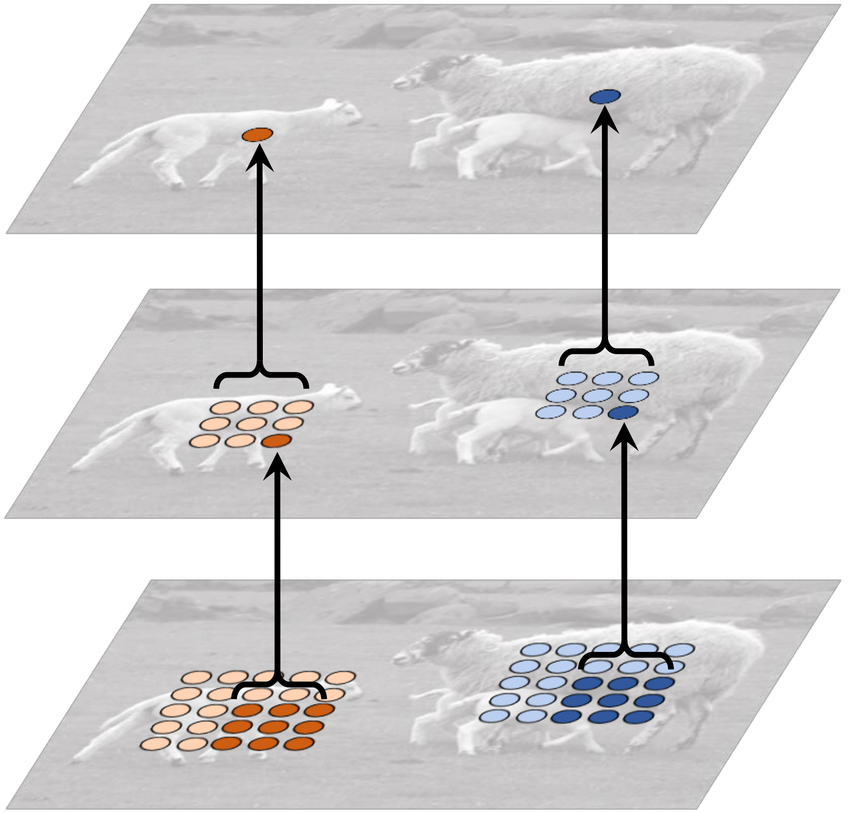
Deformable convolution:

$$y'(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + g(p_n))$$

Dai et al., “Deformable Convolutional Networks” ICCV 2017.

Why is it useful?

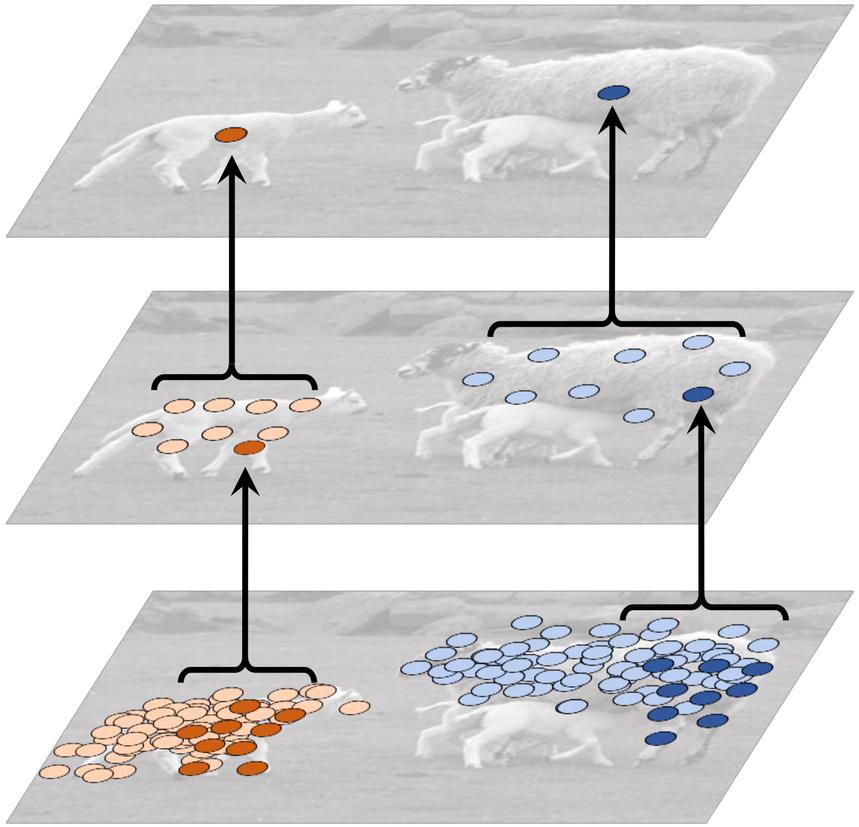
- Consider non-rigid object recognition.
  - Regular convolution:
- Problem: When an object deforms, we need another kernel to work in that pose too.



Dai et al., “Deformable Convolutional Networks” ICCV 2017.

Why is it useful?

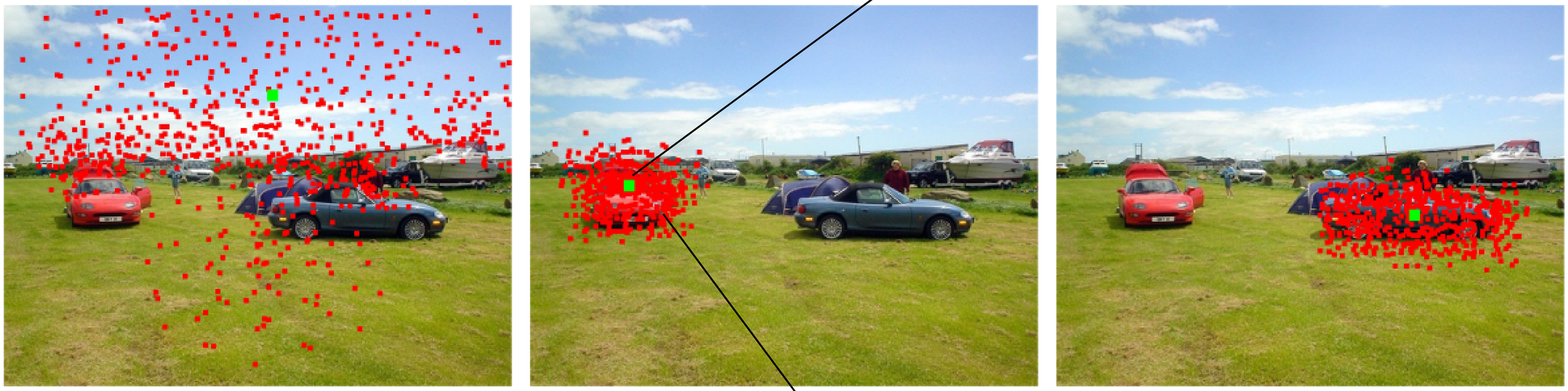
- Consider non-rigid object recognition.
  - Regular convolution:
- Problem: When an object deforms, we need another kernel to work in that pose too.
- … or deformable convolution:



Dai et al., “Deformable Convolutional Networks” ICCV 2017.

Why is it useful?

- We can inspect receptive field:



Dai et al., “Deformable Convolutional Networks” ICCV 2017.

UPSNet

- Improved accuracy over Panoptic FPN
  - joint things and stuff training via panoptic head
  - deformable convolutions
- Not very efficient (4-6 FPS): why?
  - RoI pooling
- How about using proposal-free instance segmentation?

Overview

- Kirillov et al., “Panoptic Feature Pyramid Networks”, CVPR 2019.
- Xiong et al., “UPSNet: A Unified Panoptic Segmentation Network”, CVPR 2019.
- **Cheng et al., “Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation”, CVPR 2020.**
- Li et al., “Fully Convolutional Networks for Panoptic Segmentation”, CVPR 2021.

Hough Transform

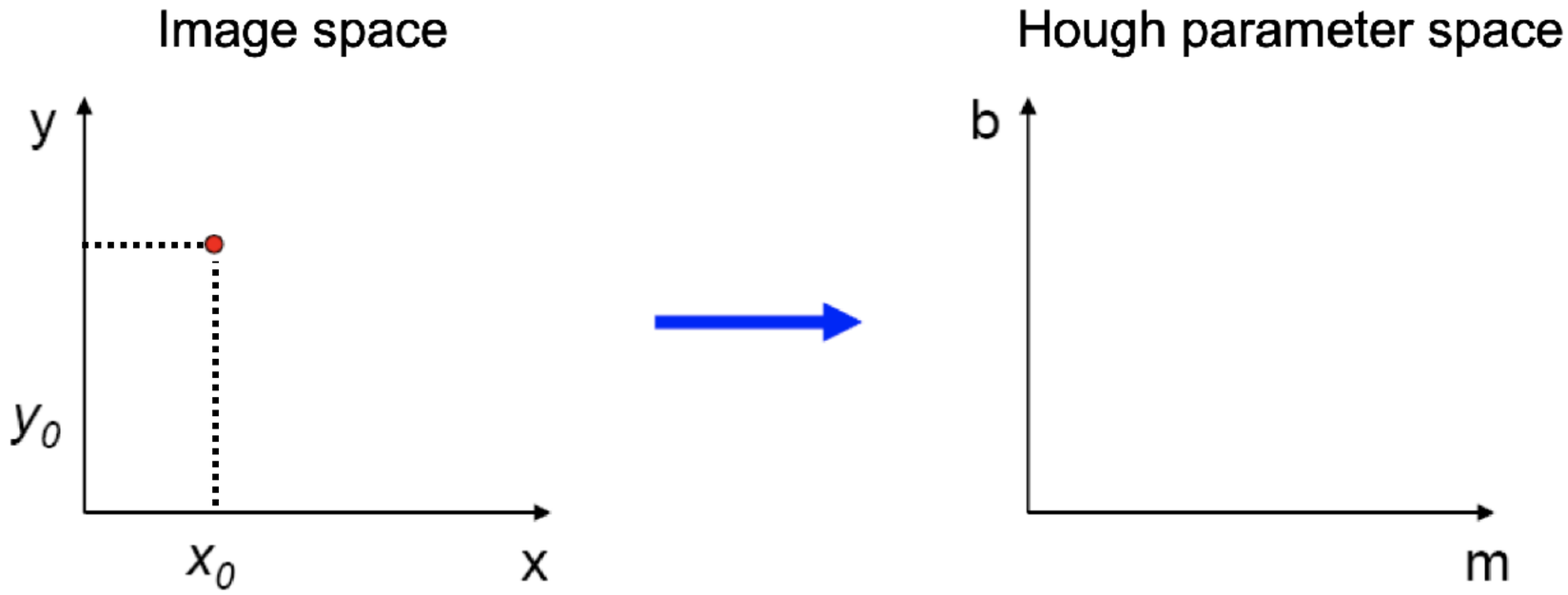
- Before deep learning dominance: detection-as-voting.

Hough voting

- Detect analytical shapes (e.g., lines) as peaks in the dual parametric space
- Each pixel casts a vote in this dual space
- Detect peaks and 'back-project' them to the image space

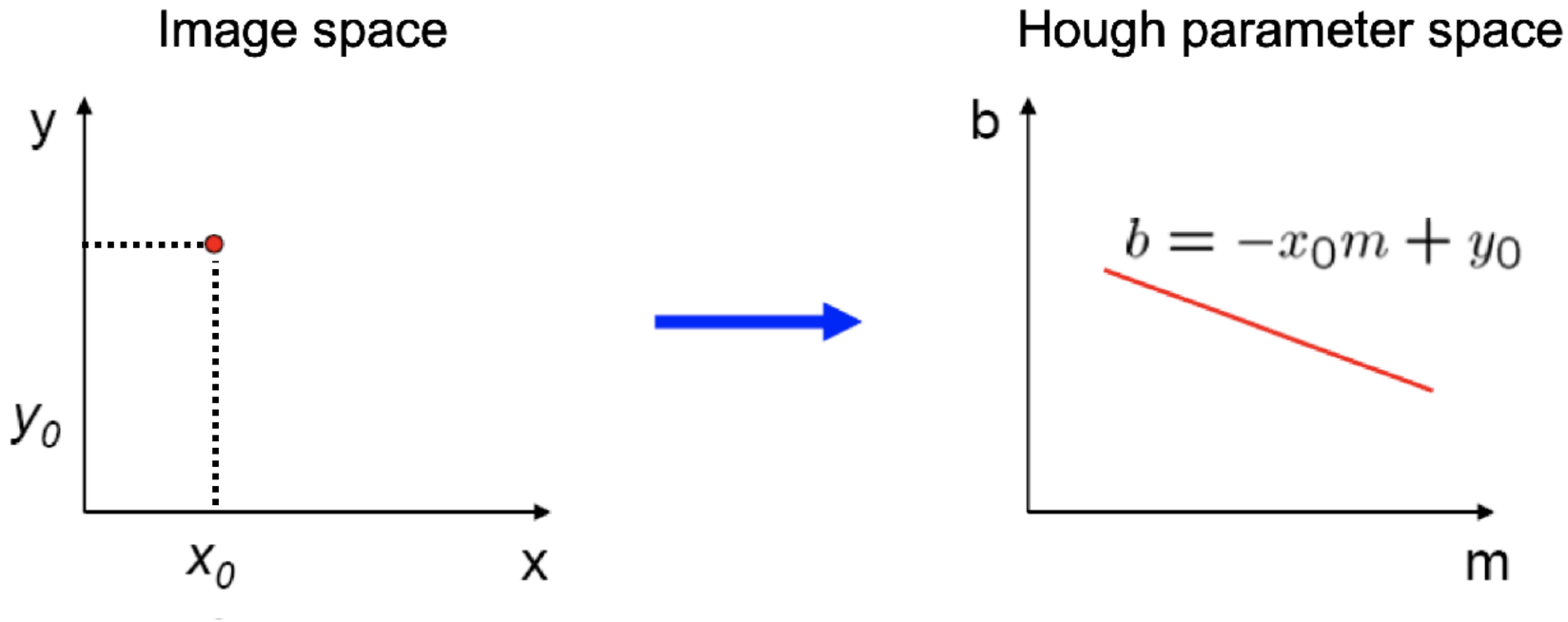
Example: Line Detection

- Each edge point in image space casts a vote



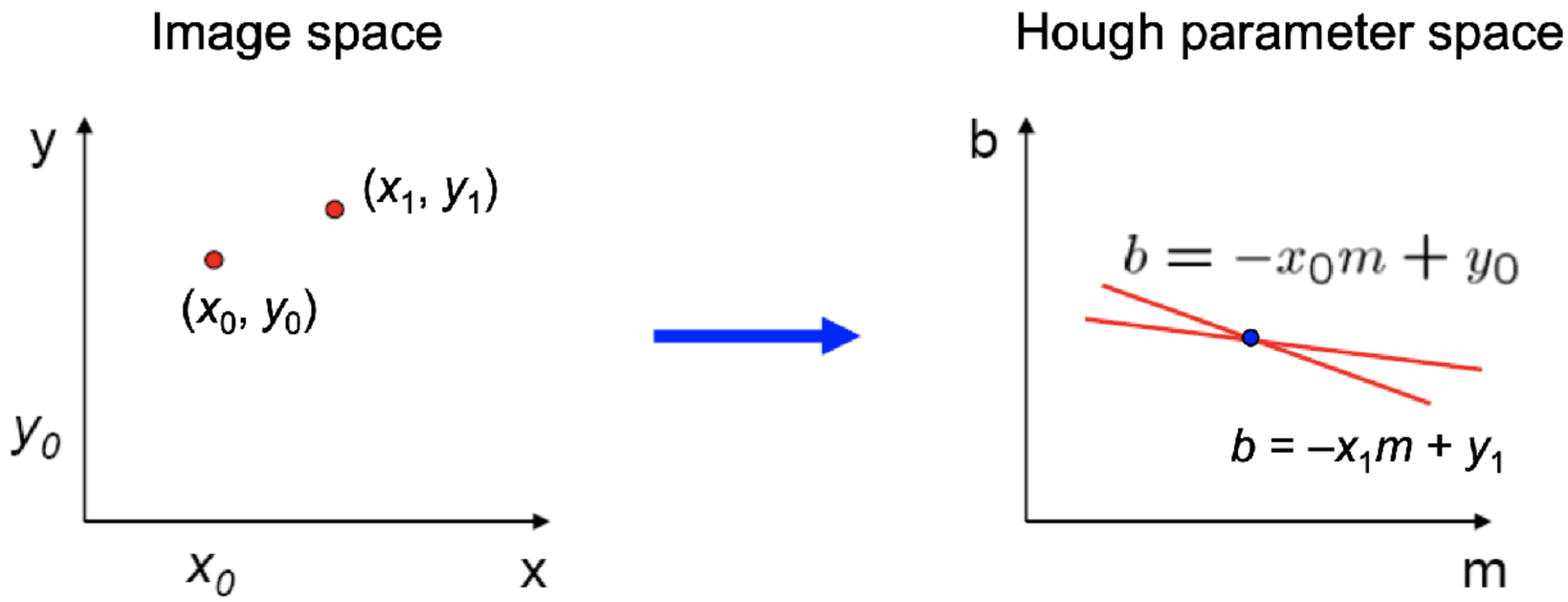
Example: Line Detection

- Each edge point in image space casts a vote
- The vote is in the form of a line that crosses the point



Example: Line Detection

- Accumulate votes from different points in (discretized) parameter space
- Read-out maxima (peaks) from the accumulator



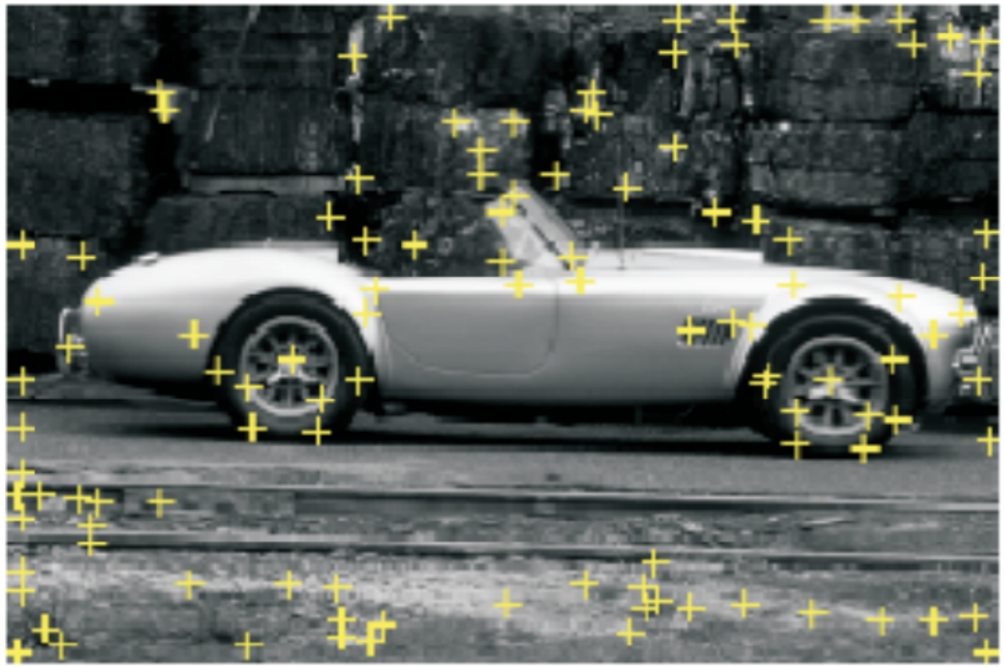
Object Detection as voting

- Idea: Objects are detected as consistent configurations of the observed parts (visual words)

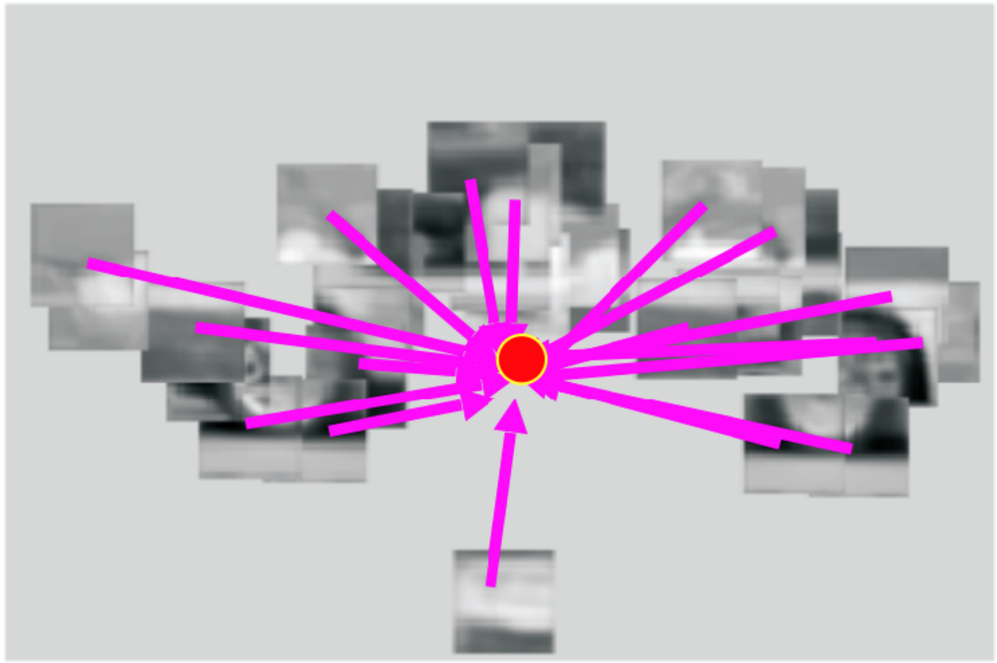
Leibe et al., Robust Object Detection with Interleaved Categorization and Segmentation, IJCV'08

Object Detection

- Training



Interest point detection
(SIFT, SURF)

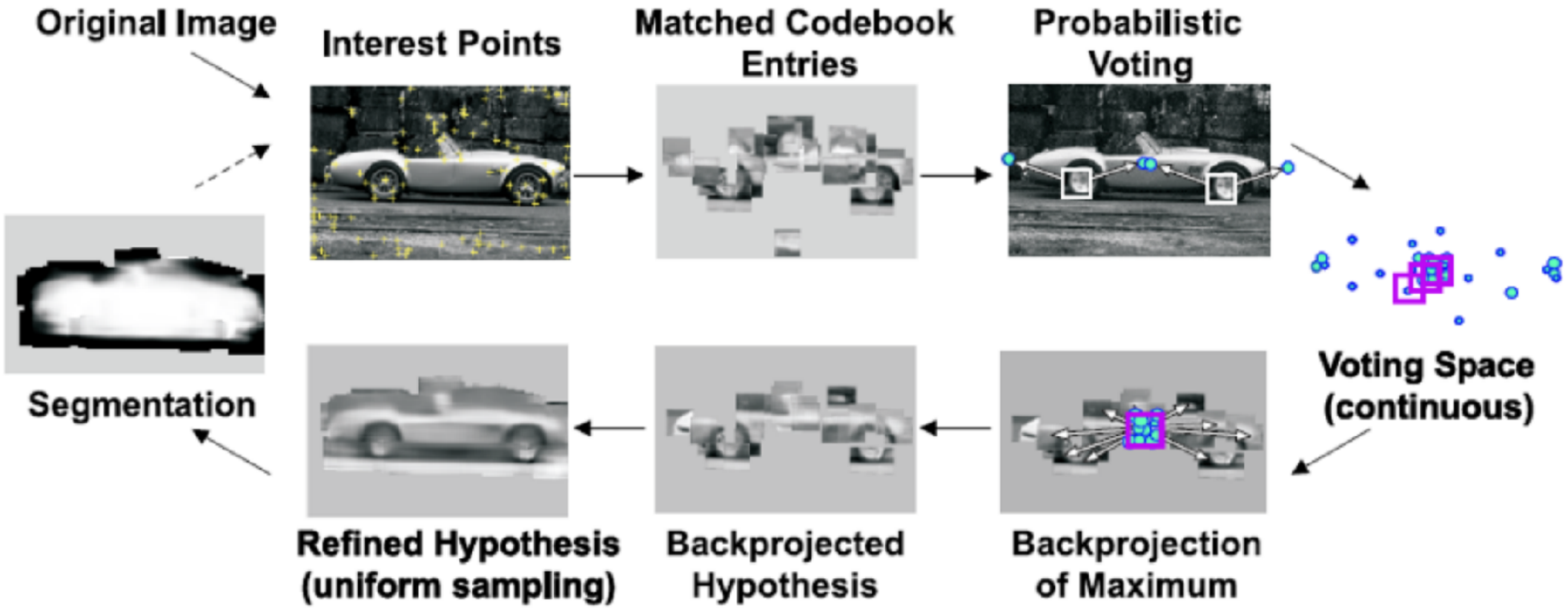


Center point voting

Leibe et al., Robust Object Detection with Interleaved Categorization and Segmentation, IJCV'08

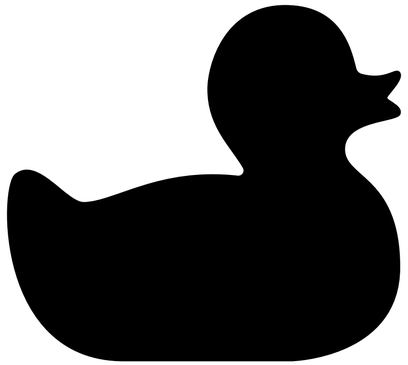
Object Detection

- Inference (test time)



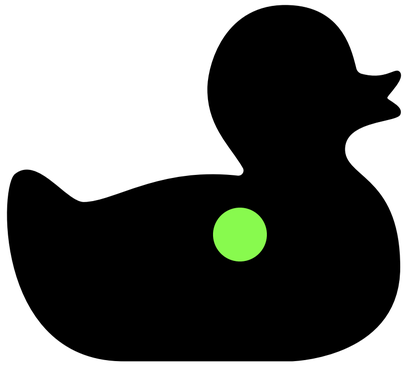
Instance mask representation

- We can apply the same concept to instance segmentation:



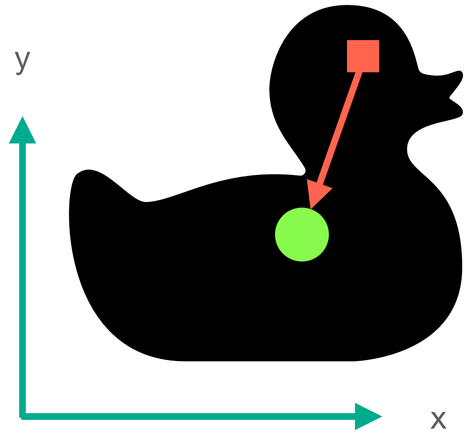
Instance mask representation

- Define the mask center:



Instance mask representation

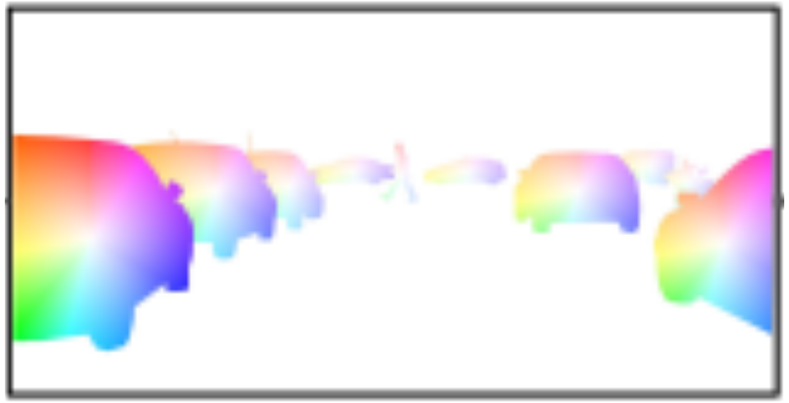
- Every pixel can be defined as an offset w.r.t. the center:



Instance mask representation

- We can learn a model predicting such offsets (regression):

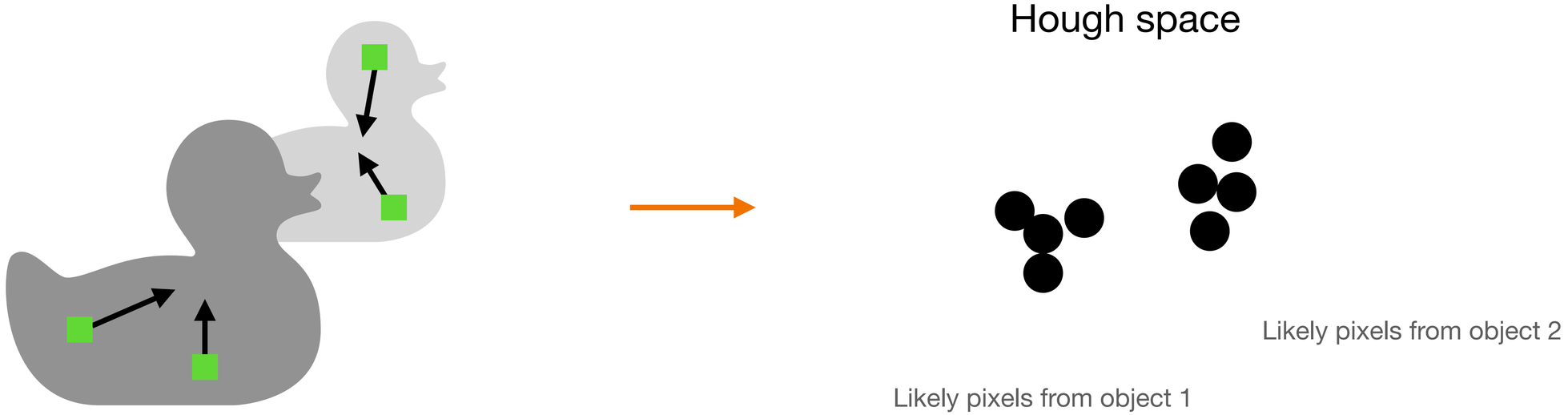
Colour-coding of the offsets:



- Alternatively, we can quantise the offsets and learn a classifier (Uhrig et al., 2016)

Instance mask representation

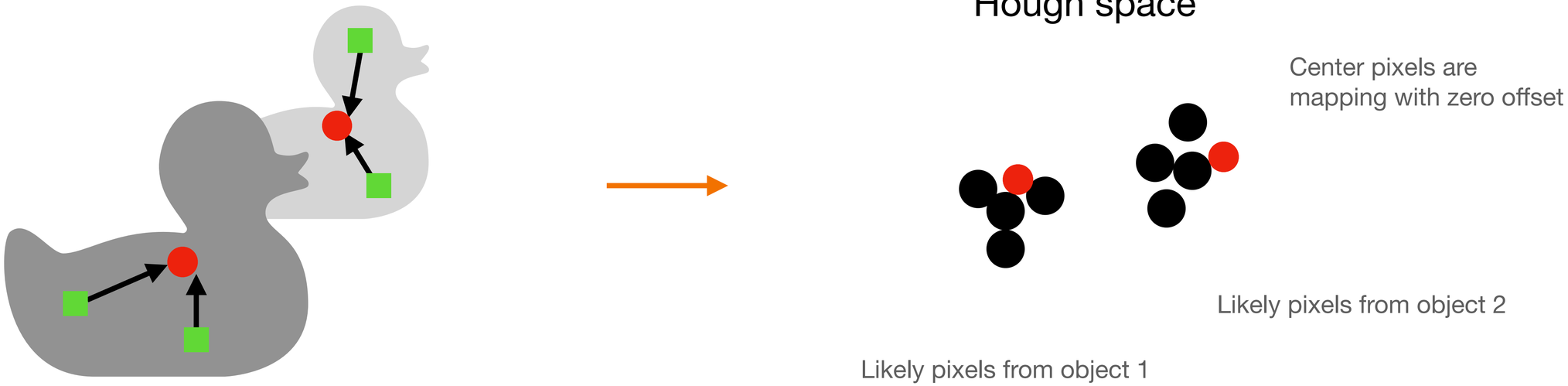
- At test time, each pixels votes for the object center to which it belongs:



- We can then cluster the pixels in the Hough space

Hough space clustering

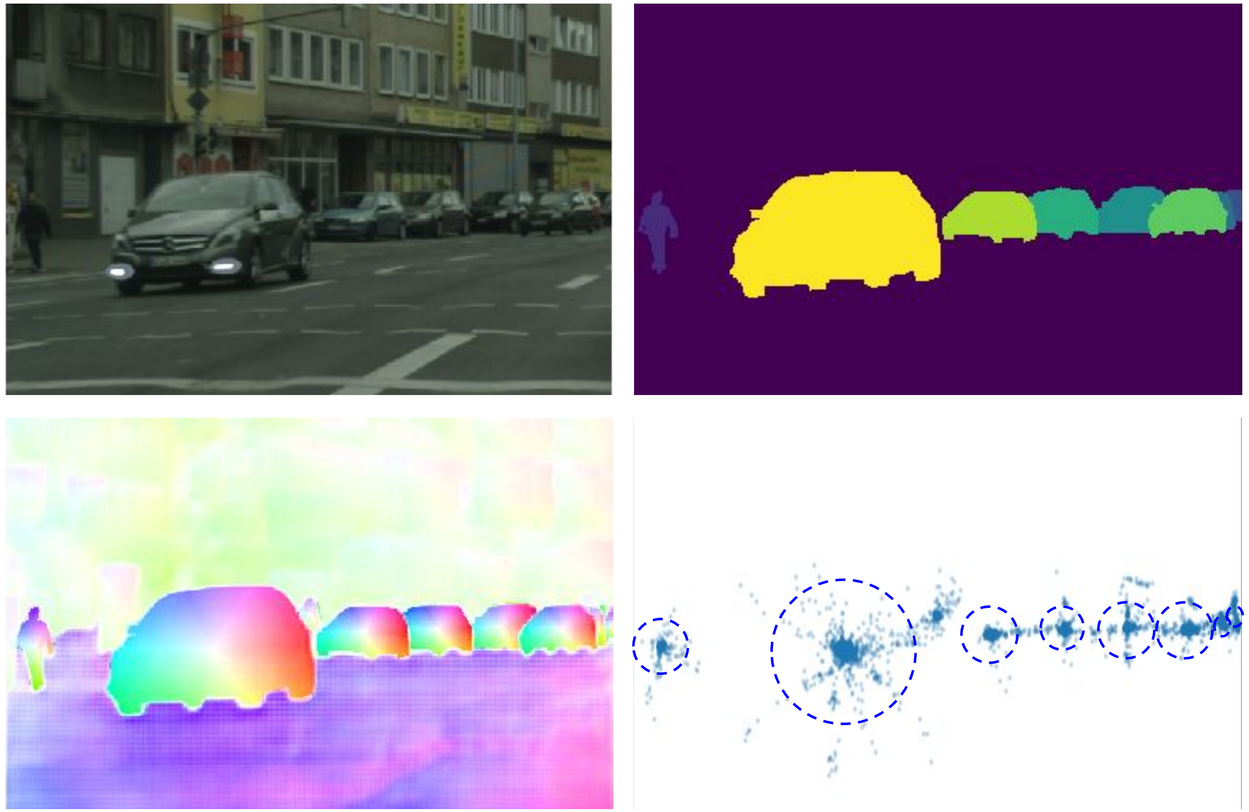
- We still need to know how many clusters.
- One approach: learn to predict centre pixels:



- Use center pixels with a fixed radius (hyperparameter) for clustering

Hough space clustering

- Real-world example (Neven et al., 2019):

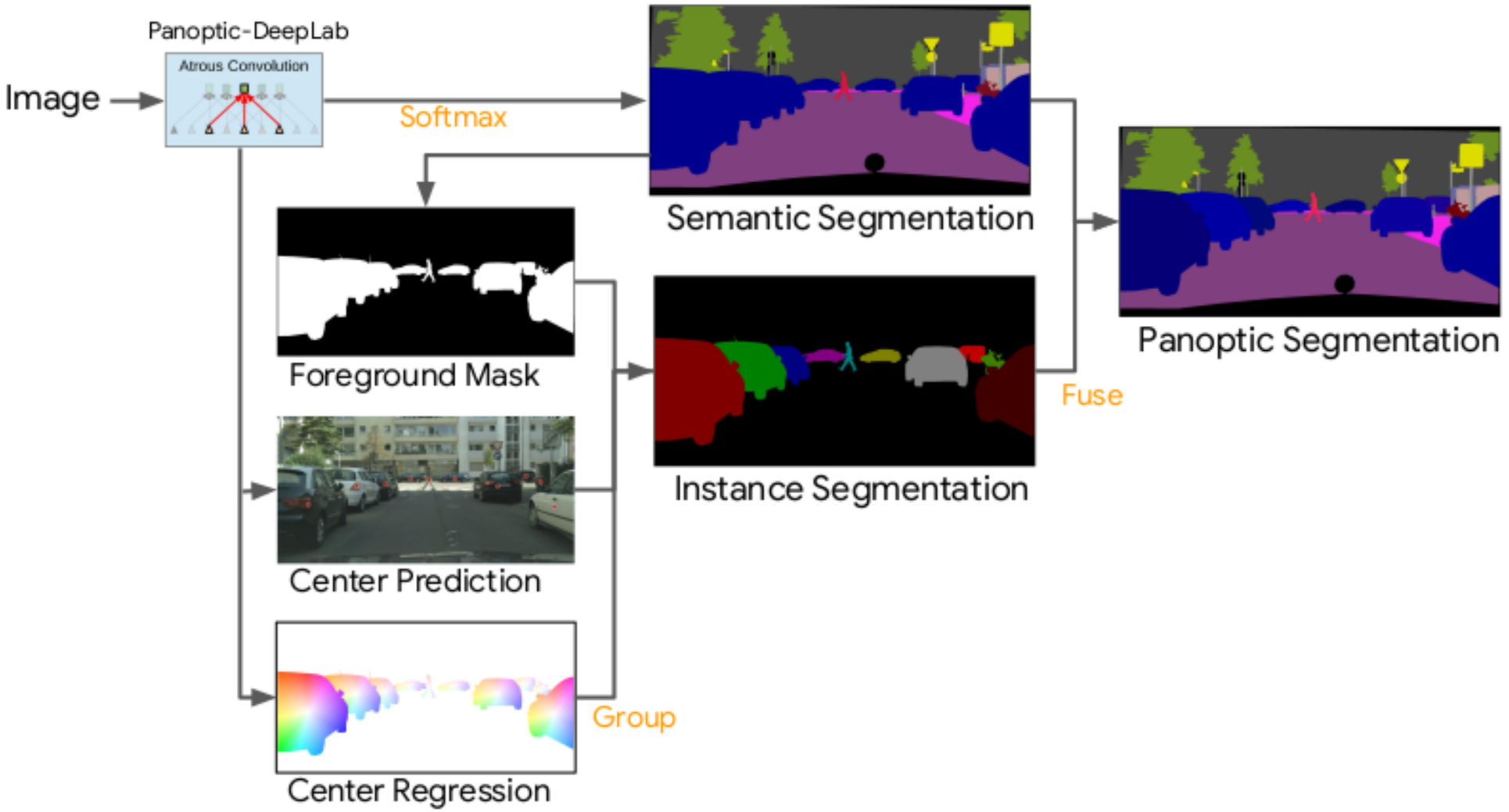


Back to panoptic segmentation

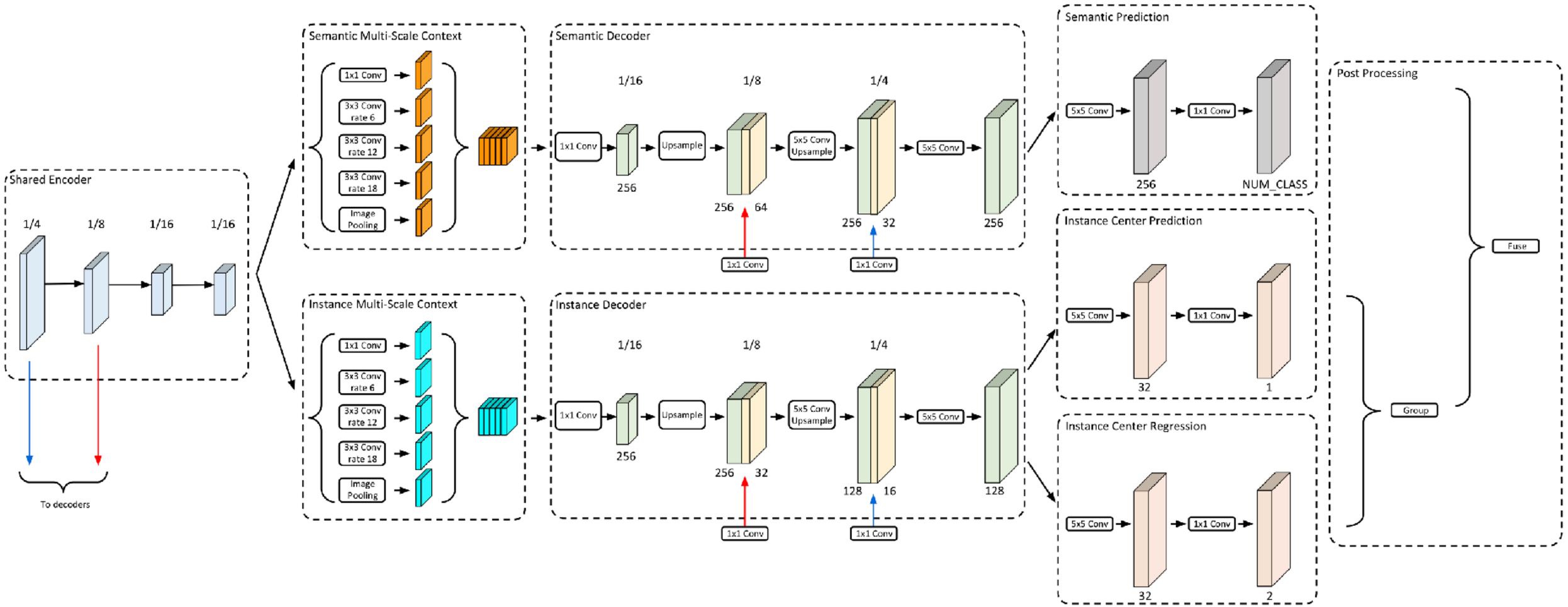
- Hough voting representation suits well panoptic segmentation:
  - no need for RoI pooling;
  - decoder architecture (not necessarily shared) can be similar between instance and semantic segmentation;
  - typically much faster to process.

- Panoptic-DeepLab

Panoptic-DeepLab

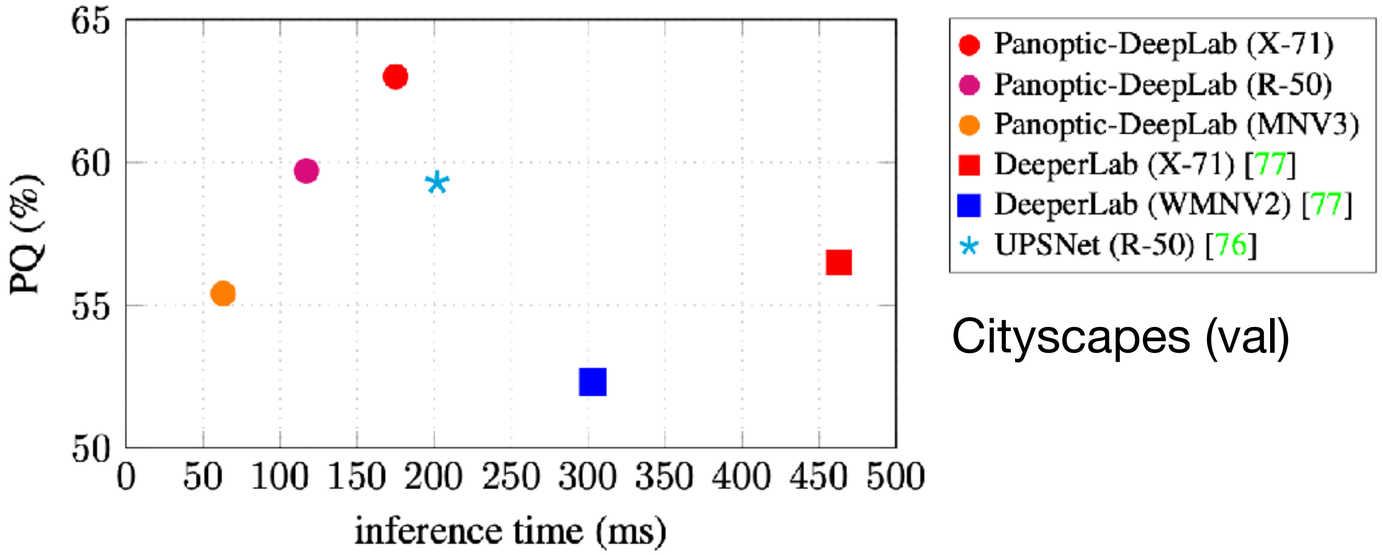


Panoptic-DeepLab



Panoptic-DeepLab

- Proposal-free panoptic segmentation approach
- Conceptually simple: the semantic and instance branches have similar fully convolutional architecture
- Competitive accuracy, yet more efficient (e.g. than UPSNet):

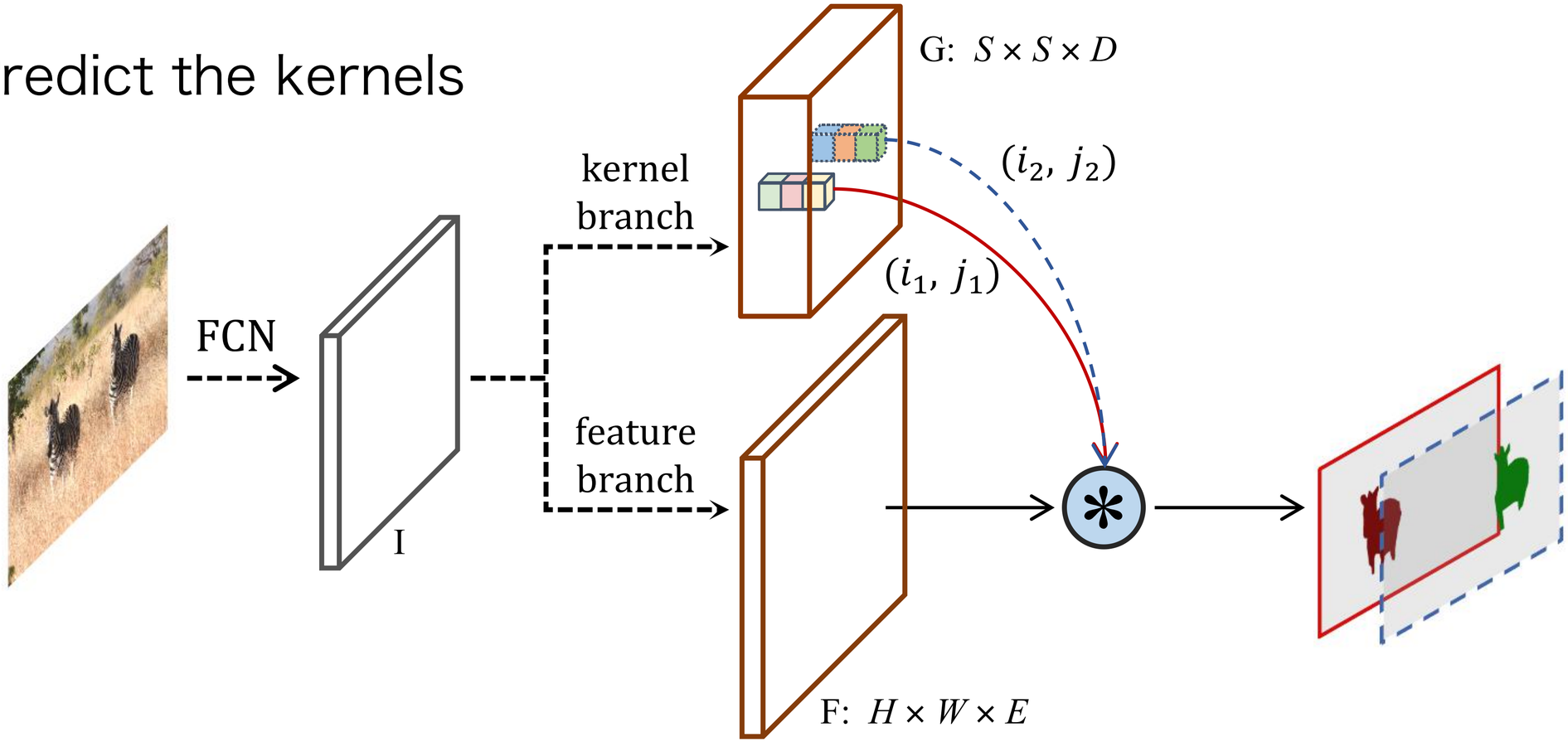


Cityscapes (val)

Even simpler?

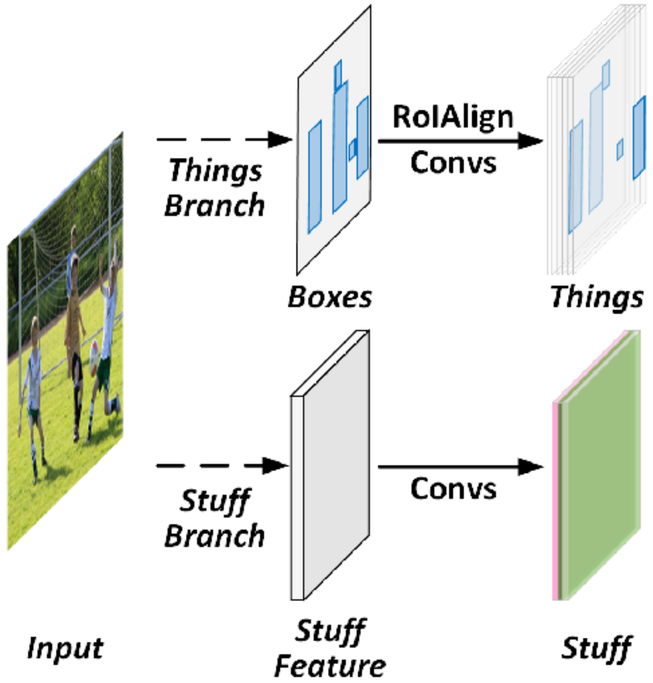
Recall SOLOv2:

Idea: Predict the kernels

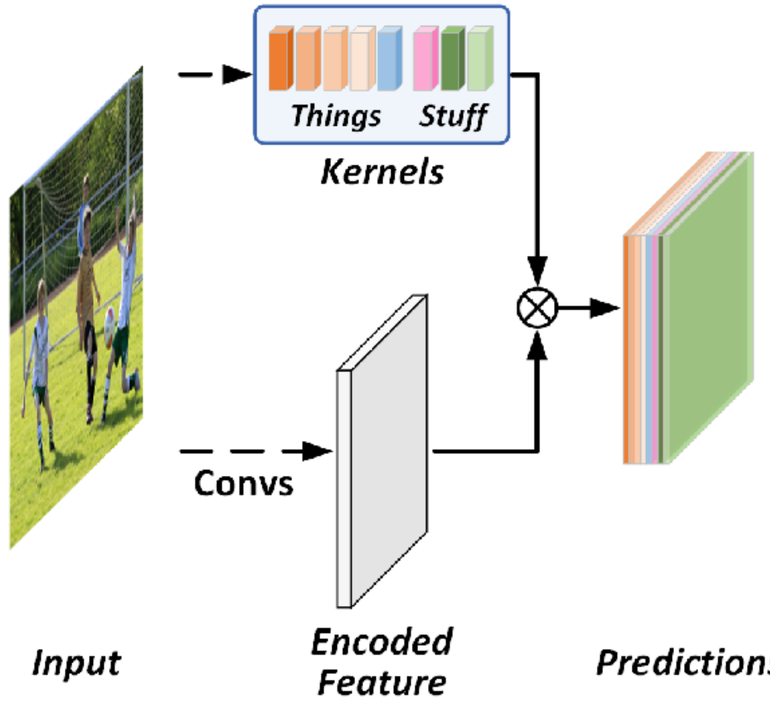


Panoptic FCN

Panoptic FPN

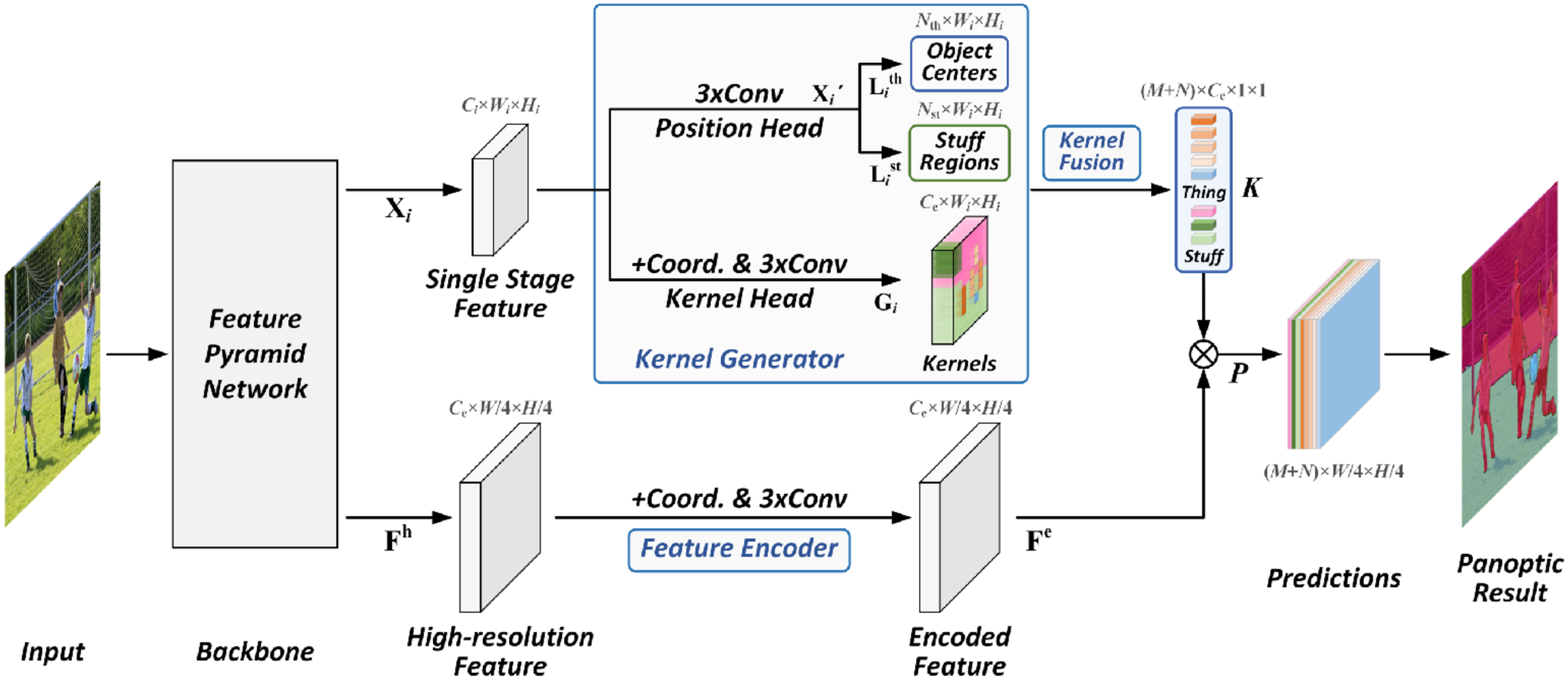


Panoptic FCN (Fully Convolutional Network)



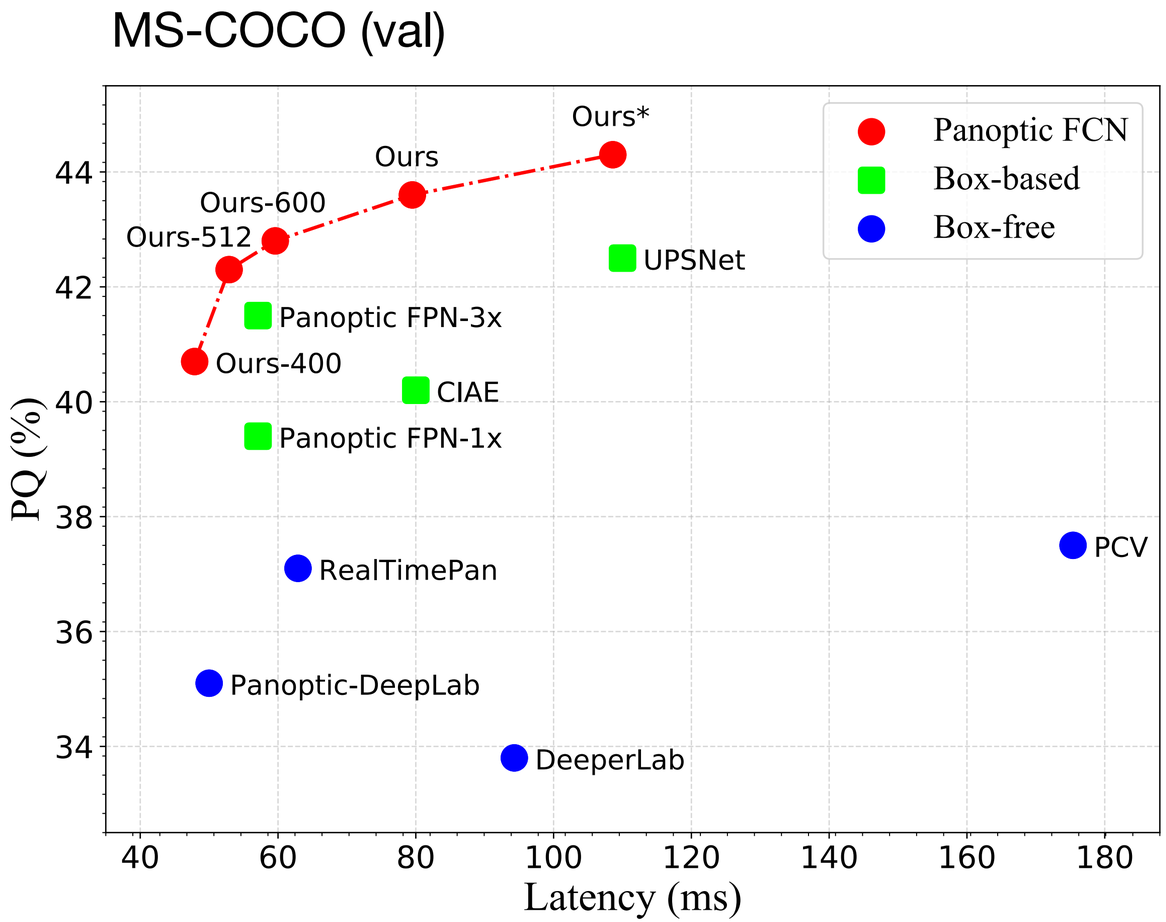
Li et al., “Fully Convolutional Networks for Panoptic Segmentation”, CVPR 2021.

Panoptic FCN



Li et al., “Fully Convolutional Networks for Panoptic Segmentation”, CVPR 2021.

Panoptic FCN



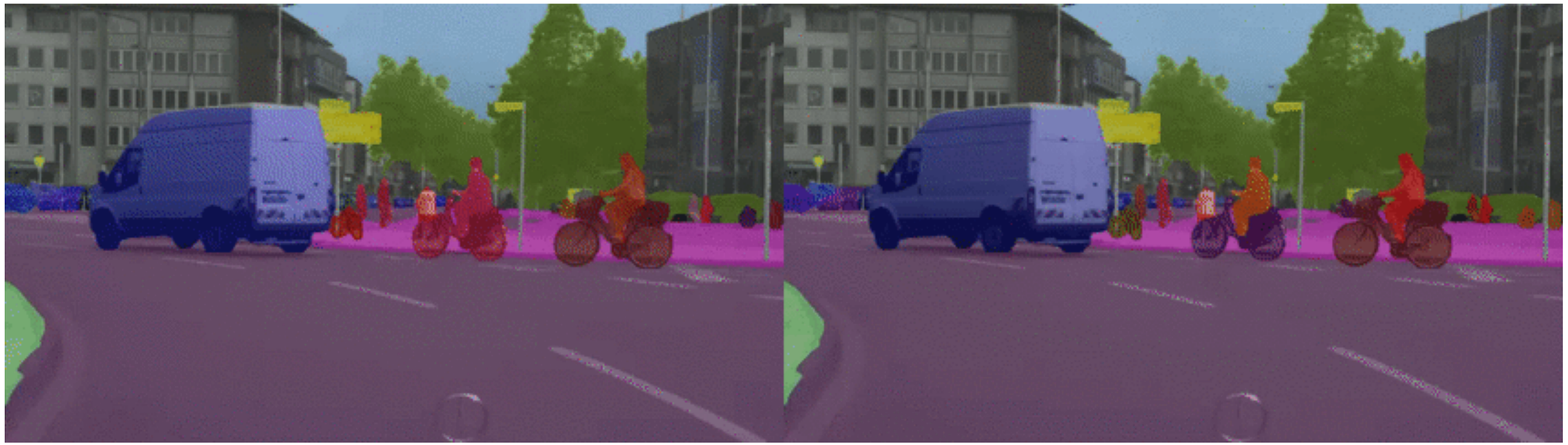
- Improved efficiency and accuracy
- Simpler architecture

Li et al., “Fully Convolutional Networks for Panoptic Segmentation”, CVPR 2021.

Panoptic FCN: Qualitative examples

Current research

Video panoptic segmentation (Kim et al., 2020):



Next lecture: Video Object Segmentation

See also:

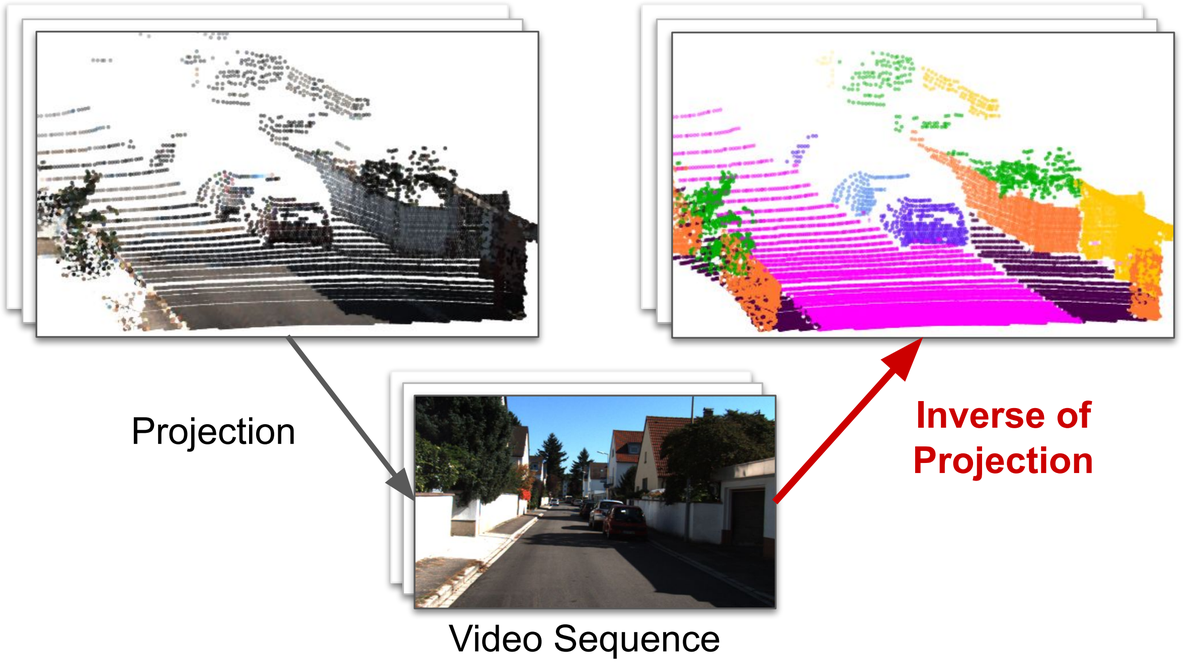
Weber et al., “STEP: Segmenting and Tracking Every Pixel” (2021).

Current research

Extending to other modalities (e.g. depth prediction):

Input: RGB video sequence

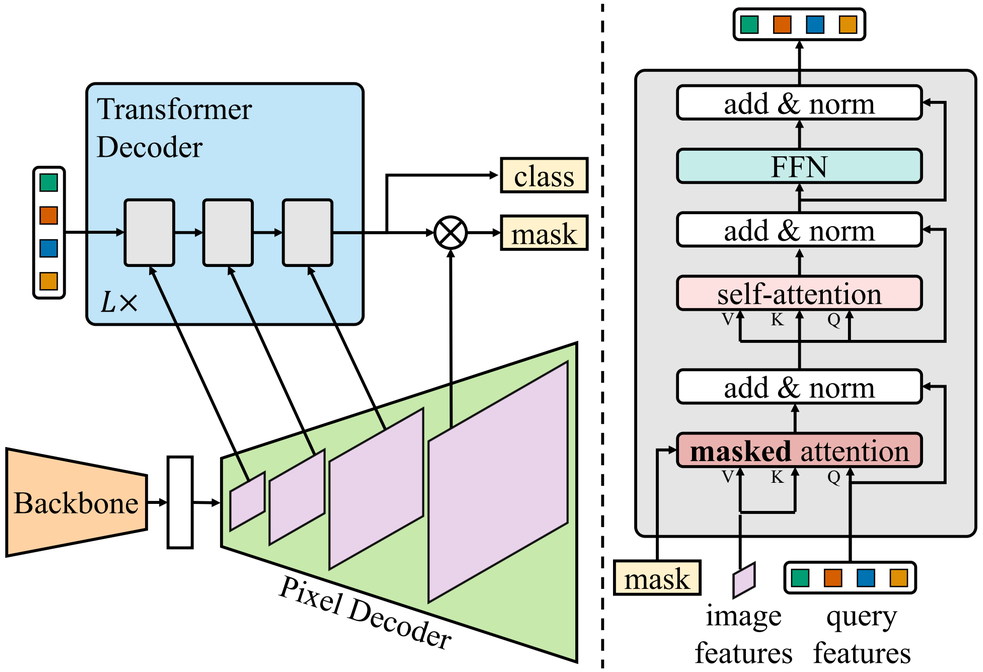
Output: semantic labels (panoptic) + depth

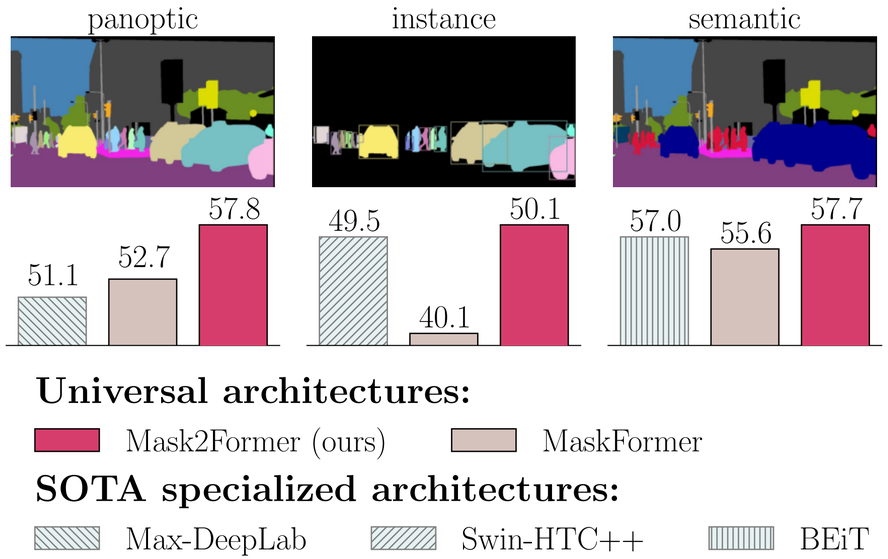


Qiao et al., “ViP-DeepLab: Learning Visual Perception with Depth-aware Video Panoptic Segmentation”, CVPR 2021.

Current research

Improved architecture with Transformers (Mask2Former):



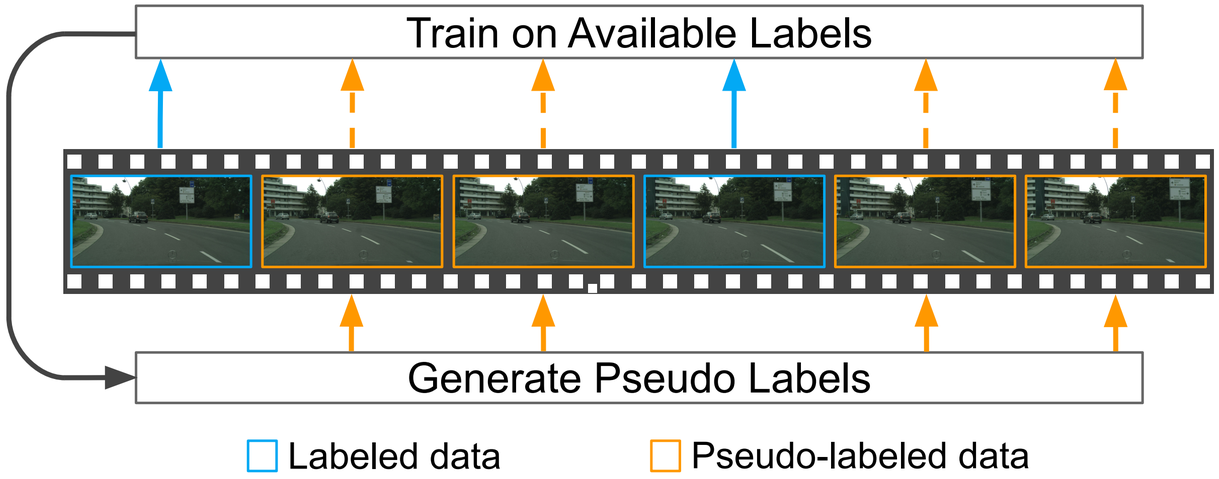


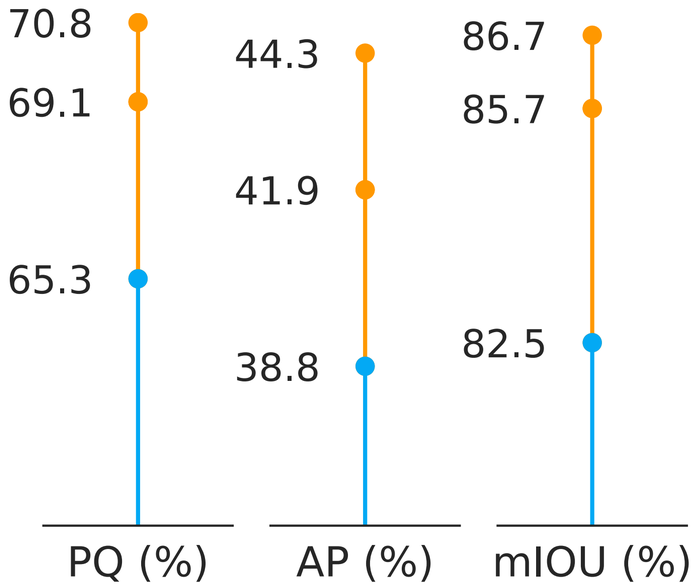
Upcoming: Transformers

Cheng et al., “Masked-attention Mask Transformer for Universal Image Segmentation”, CVPR 2022.

Current research

Semi-supervised learning:





Upcoming: Semi- and self-supervised learning

Chen et al., “Naive-Student: Leveraging Semi-Supervised Learning in Video Sequences for Urban Scene Segmentation”, ECCV 2020.

Panoptic segmentation

Ongoing research:

- Improving context reasoning (e.g. with attention layers);
- Harmonising “stuff” and “things” representation;
  - Transformers (e.g., DETR, MaskFormer) – later in the course.
- Improving efficiency (e.g. EfficientPS);
- Extending to video data (video object segmentation – next lecture);
- Adding other modalities, such as depth prediction.

Round-up

Reminders:

- Ongoing course evaluation (deadline: this Friday, 23.12);
- Exam registration (deadline: 15.01)
- Exercise 3 release tomorrow.
- Next lecture: January 10, 2023

Happy holidays!