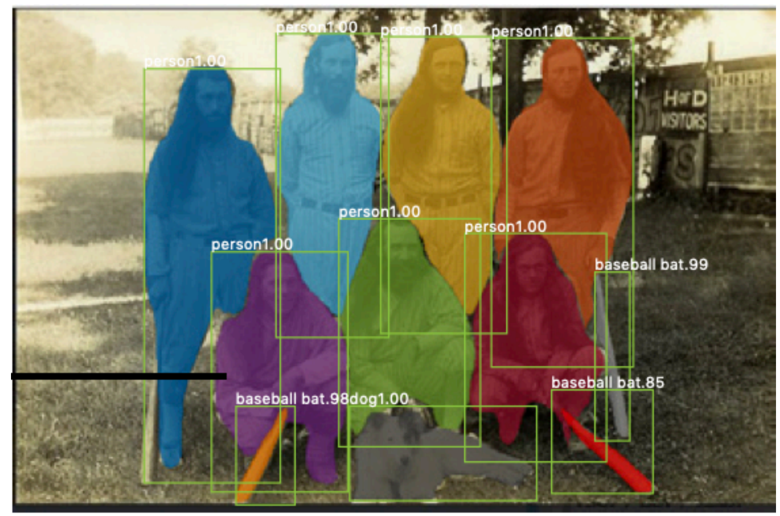
Pixel Consensus Voting for Panoptic Segmentation

Haochen Wang Ruotian Luo Michael Maire Greg Shakhnarovich

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

THE UNIVERSITY OF CHICAGO

Object detection/Instance segmentation

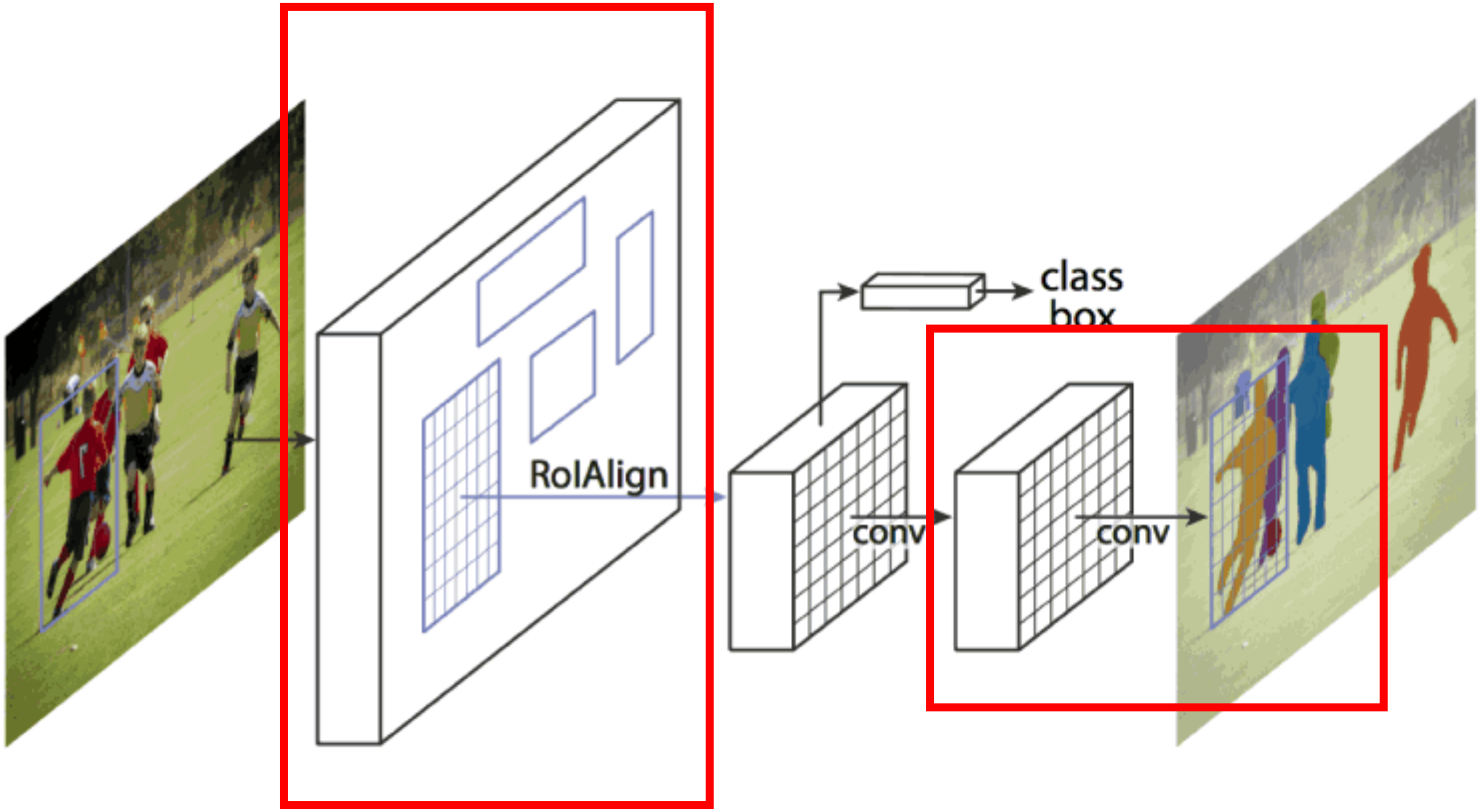


PASCAL VOC 2007 Bounding Box Detection

MSCOCO (2014): Bounding Box and Instance Mask

Object detection/Instance segmentation

- reason about densely enumerated bounding boxes, and then refine to instance masks.

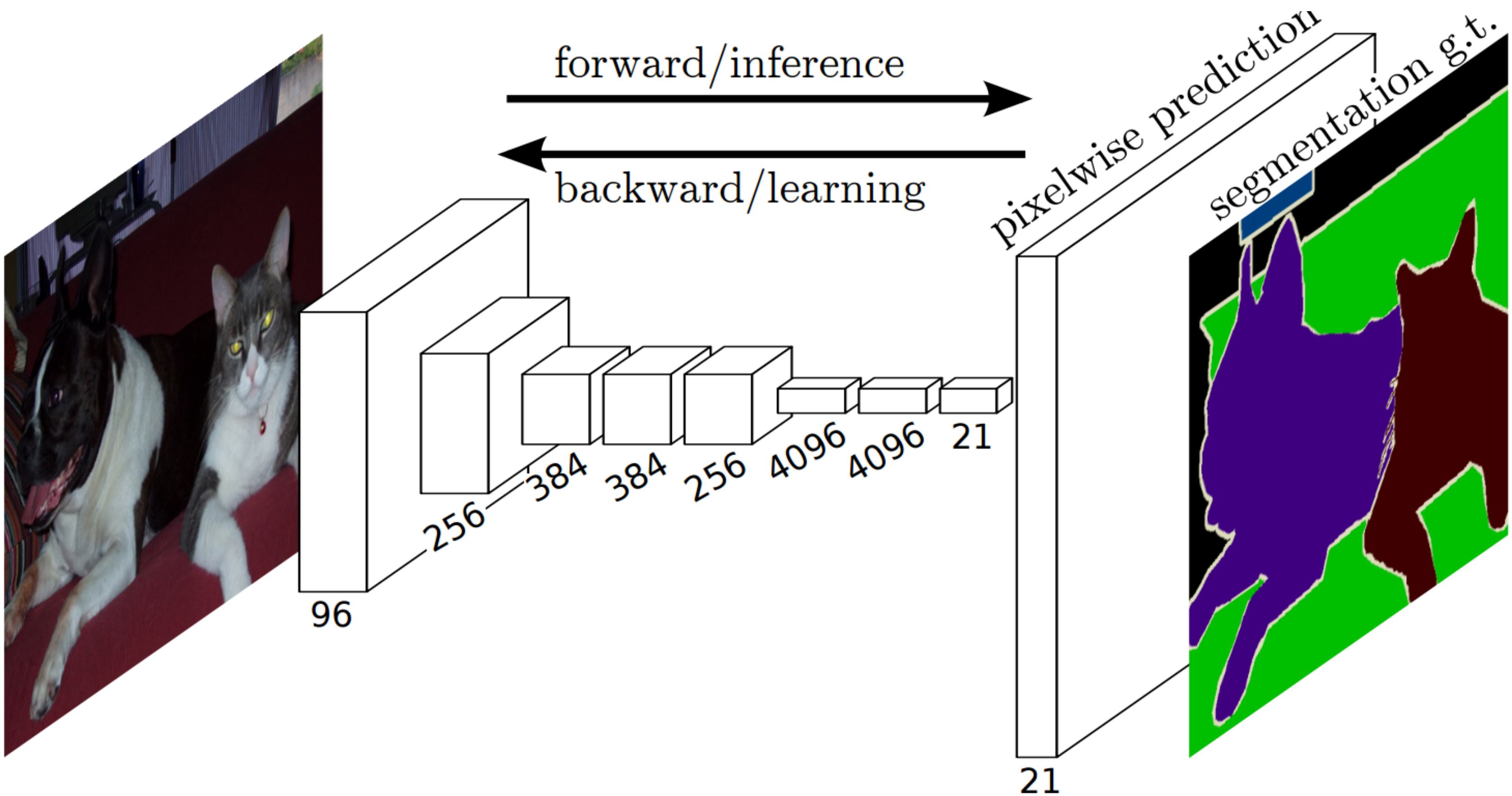


The Mask R-CNN framework for instance segmentation

Semantic segmentation

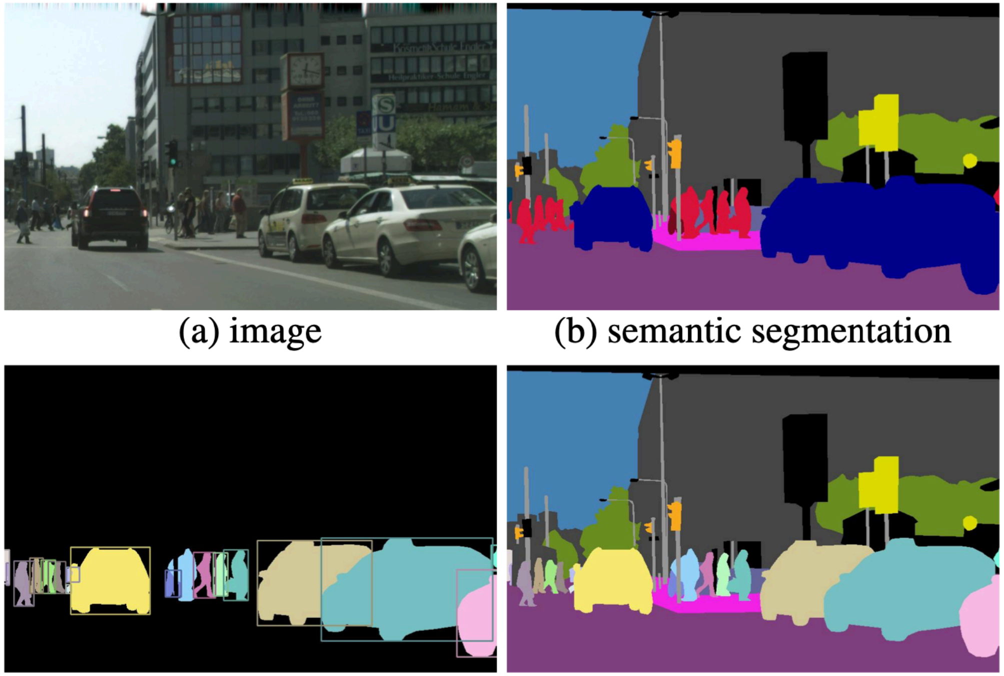
- Label each pixel in the image with a category label
- Don't differentiate instances, only care about pixels

Semantic segmentation



Panoptic Segmentation

- Instance-aware semantic segmentation.
- Things *(bicycle, dog, car, person)*
- Stuff *(pavement, ground, dirt, wall)*



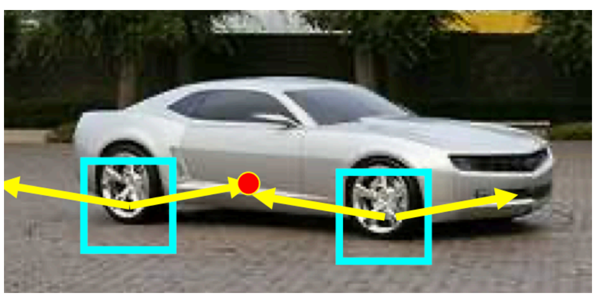
(a) image (b) semantic segmentation (c) instance segmentation (d) panoptic segmentation

Existing methods for panoptic segmentation

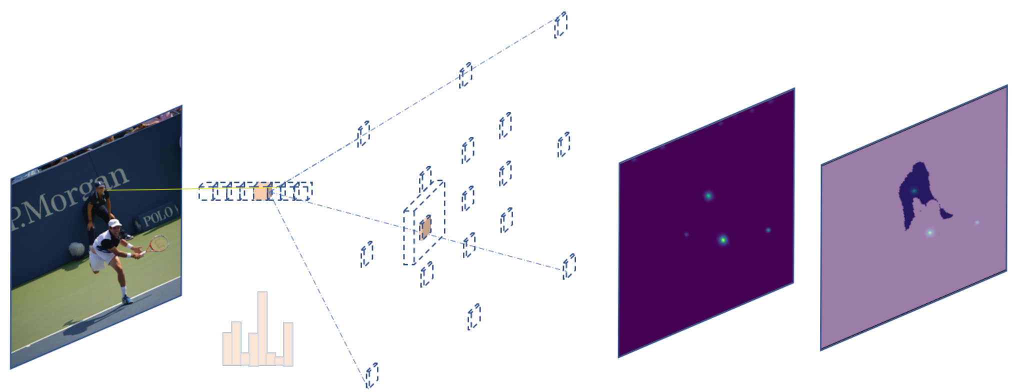
- Merge instance segmentation and semantic segmentation

However?

- Panoptic Segmentation removes the concept of boxes and focuses on pixels.
- Bounding boxes may not be the optimal intermediate representations to predict.
- Pixel Consensus Voting: pixels vote for object centroid
  - Training: per pixel classification
  - Inference: Generalized Hough Transform similar to Implicit Shape Model (ISM)



ISM

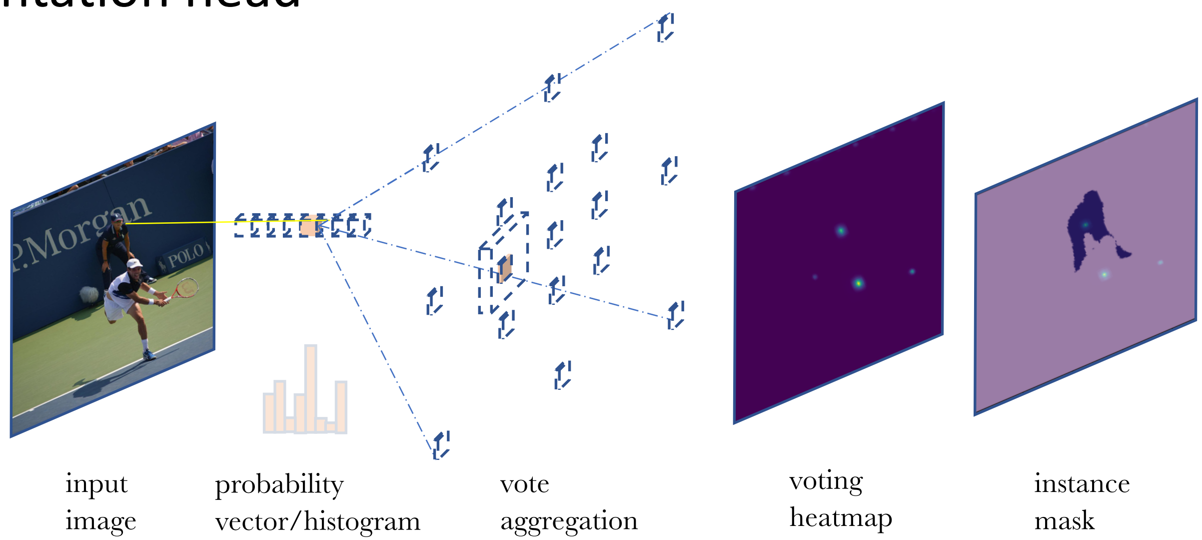


PCV

B. Leibe, A. Leonardis, and B. Schiele, Robust Object Detection with Interleaved Categorization and Segmentation, International Journal of Computer Vision, Vol. 77(1-3), 2008.

Pipeline

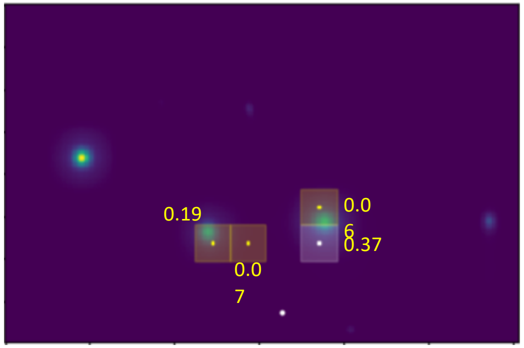
- Discretize regions around each pixel. CNN classifies the likely regions that contain centroid
- Vote aggregation: probabilities at each pixel are cast to corresponding regions through dilated deconvolution.
- Peaks in voting heatmap are detections.
- Back-projection by convolving the query filter within a peak region to get an instance mask.
- Category information provided by the parallel semantic segmentation head



Classification or Regression

- A 2d offset vector is a limited representation.
- Regression fails to capture uncertainty. Spurious peaks and false positives.
- Insight echoed by bounding box detectors: classify a proposal into anchors rather than direct regression.

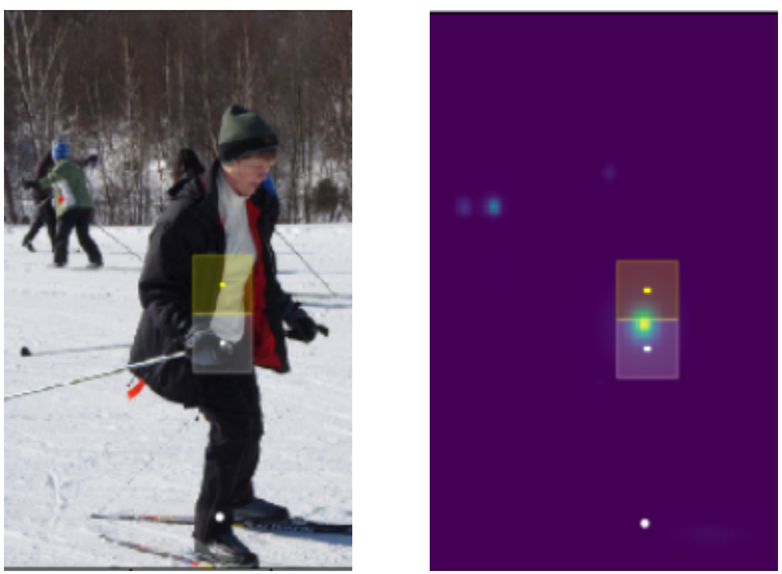
Direct regression might create spurious peaks



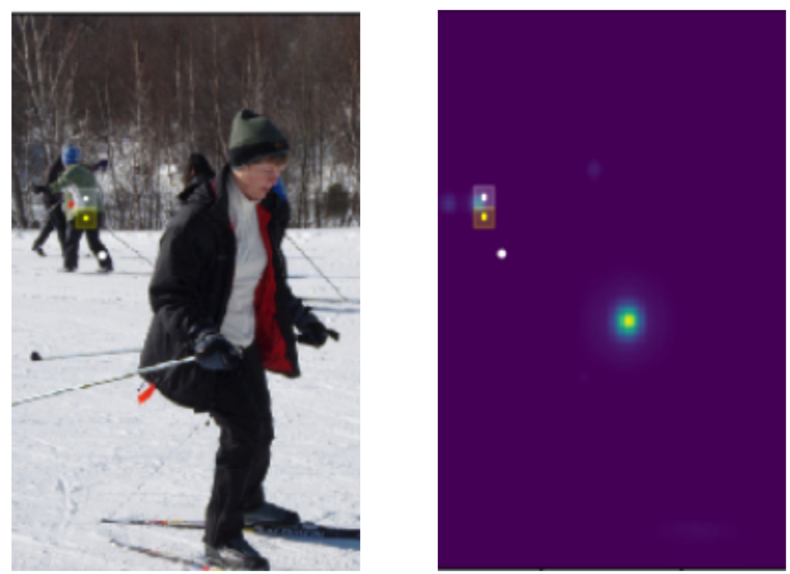
Classification models uncertainty about regions

Classification or Regression

- Perfect centroid prediction is unnecessary. What matters is consensus.
- Tolerance for coarse prediction depends on the scale.
- Easier to learn and train.



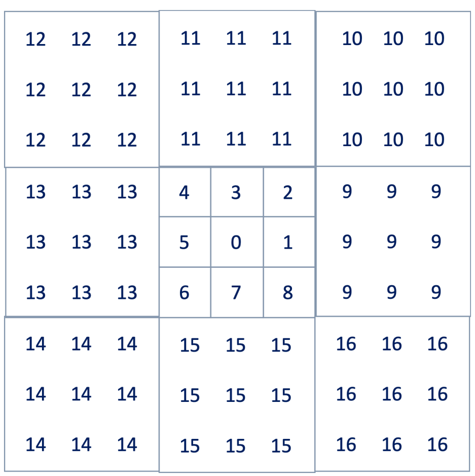
Pixels of a large object need only a rough estimate



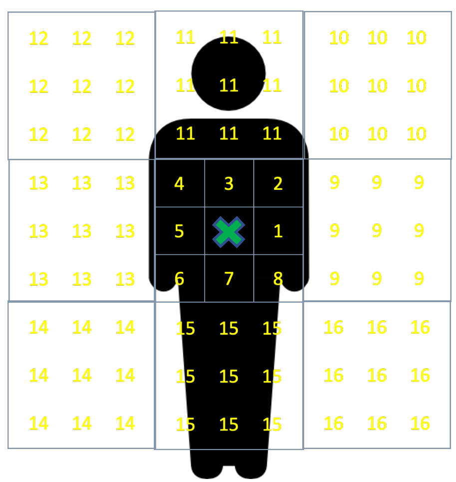
Pixels of a small object need to be precise

Discretization Scheme

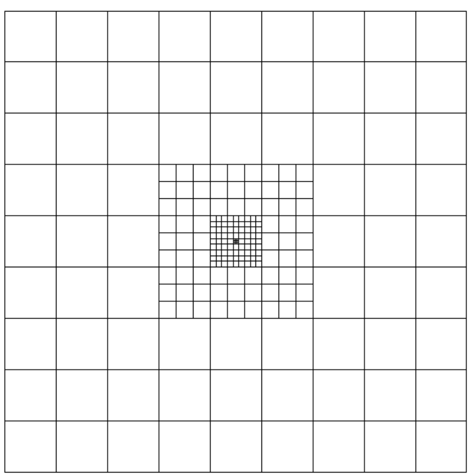
- Discretize regions around each pixel into radially expanding cells.
- **Voting mask** records the ground truth label if the centroid falls into a cell.
- Full discretization has 233 cells covering 243^2 regions at 1/4 input resolution.



A toy discretization of the 9 x 9 region around location (4, 4)
There are 17 cells.



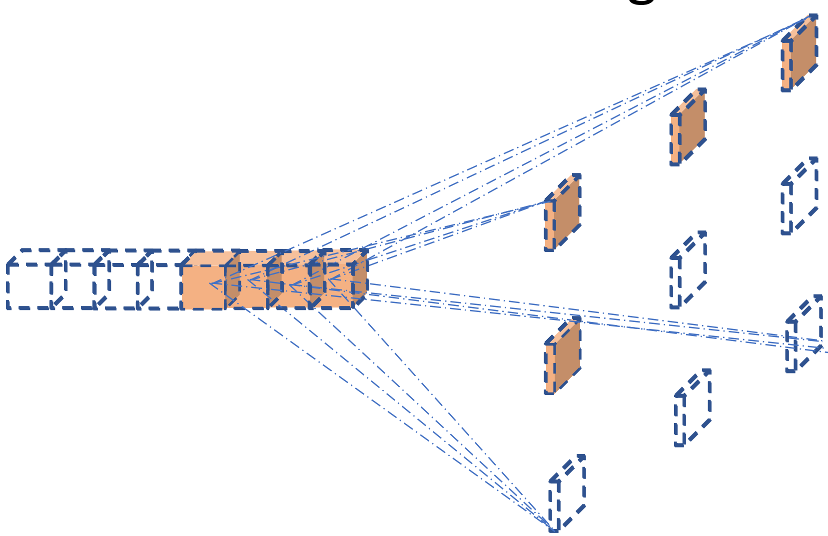
To assign voting label, align mask center with the current pixel, read off the label from the region that contains the centroid.



Full discretization of the 243^2 region around each pixe
There are 233 cells.

Voting as Dilated Deconvolution

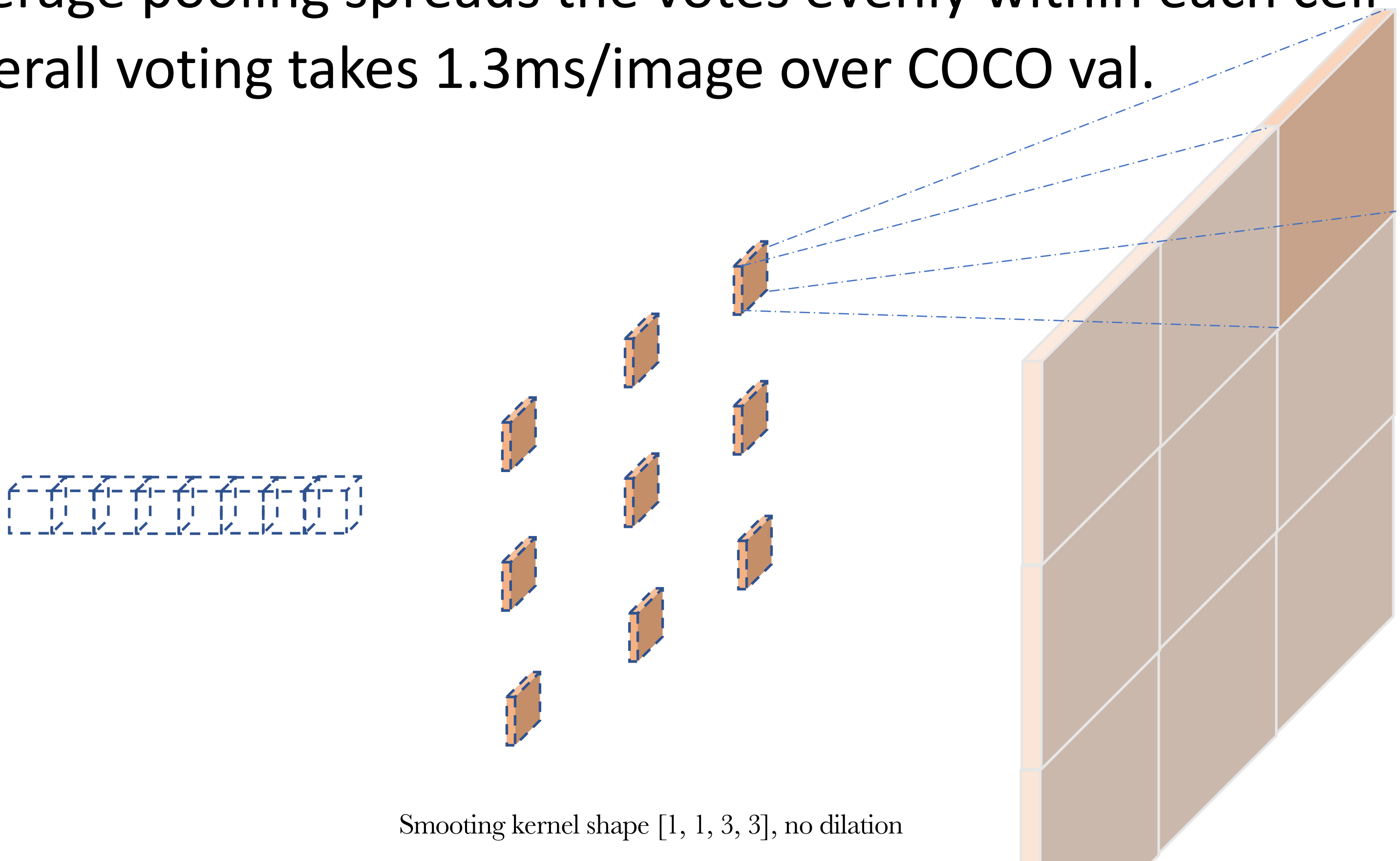
- Spreads probabilistic votes from a point to spatial locations
- Dilation allows a pixel to send its votes afar
- Fixed kernel parameters to enable voting



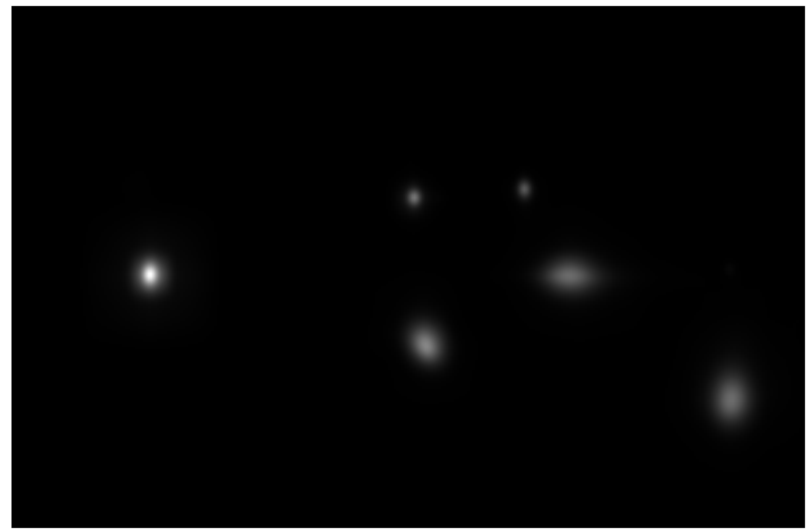
Voting kernel shape [9, 1, 3, 3], dilation 3

Voting as Dilated Deconvolution

- The previous step sends votes to points, but we are voting for regions.
- Average pooling spreads the votes evenly within each cell
- Overall voting takes 1.3ms/image over COCO val.

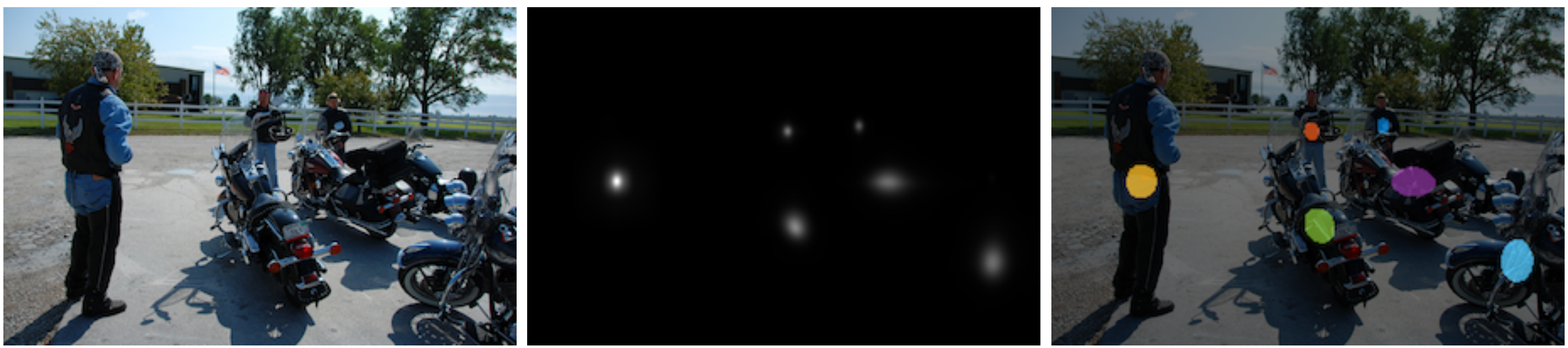


Smooting kernel shape [1, 1, 3, 3], no dilation

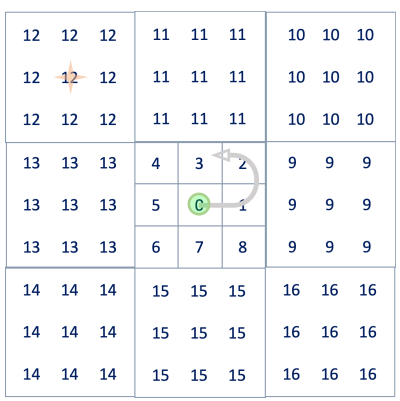
Peak detection

- Thresholding + connected component

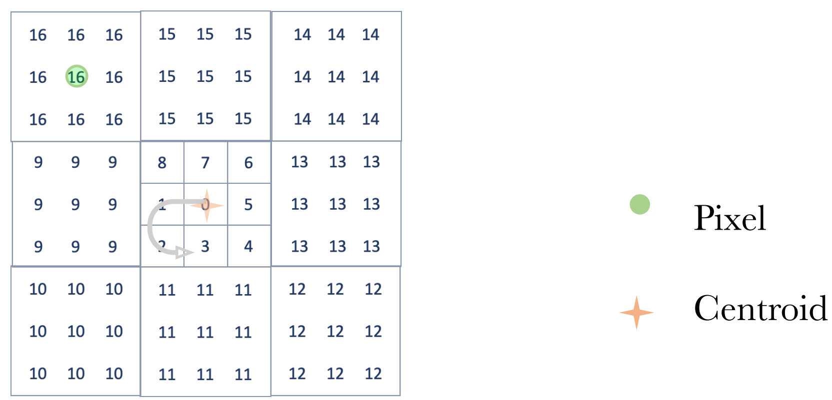


Back-projection as convolutional filtering

- For a peak, find the pixels that favor this peak above all others
- Query filter: "inward reflection" of the voting mask. Convolve in a peak region to pick up instance mask.
- 81.8ms/image.

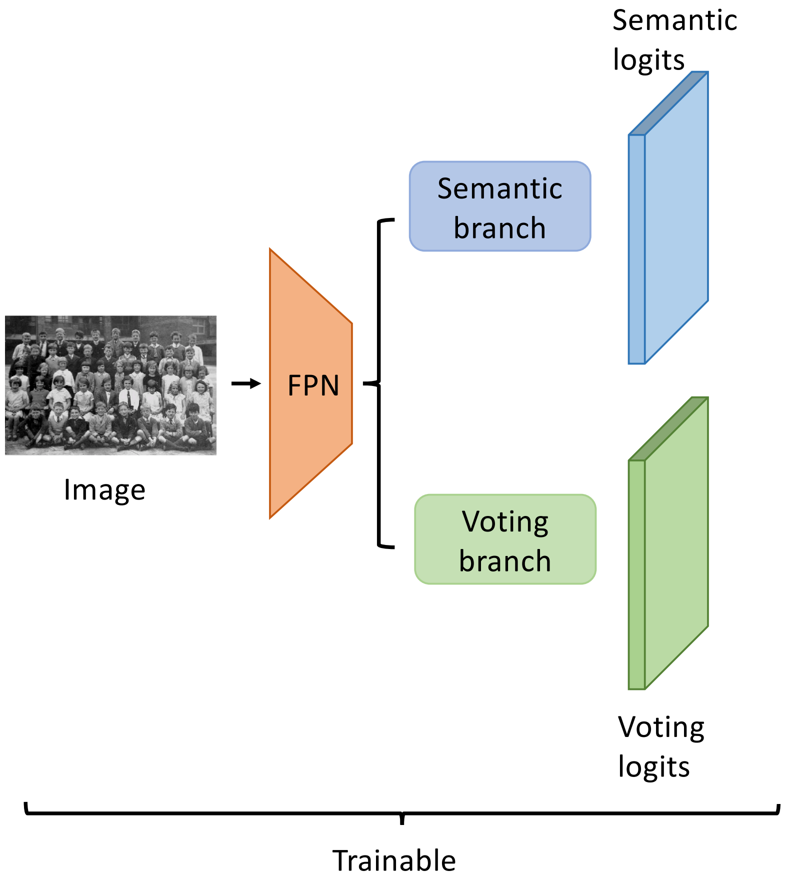


Voting mask: gt votes by a single pixel for possible centroids



Query filter: gt votes by surrounding pixels for a particular centroid

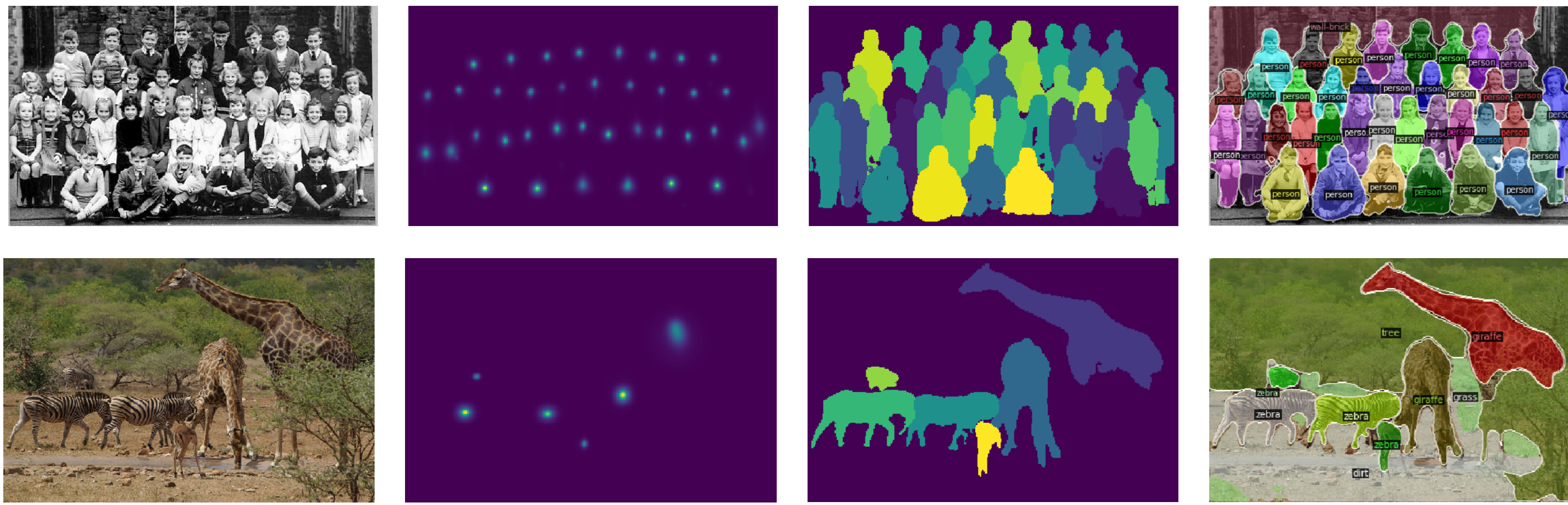
Network


Semantic logits
Semantic branch
FPN
Image
Voting branch
Voting logits
Trainable

Qualitative results

	Methods	Split	PQ	SQ	RQ	PQ^{th}	SQ^{th}	RQ^{th}	PQ^{st}
Mask R-CNN	PFPN [24] (1x)	val	39.39	77.83	48.31	45.91	80.85	55.35	29.55
	PFPN [24] (3x)	val	41.48	79.08	50.52	48.26	82.21	57.85	31.24
	UPSNet [62] (1x)	val	42.5	78.0	52.5	48.6	79.4	59.6	33.4
Single Stage Detection	SSPS [59]	val	32.4	-	-	34.8	-	-	28.6
	SSPS [59]	test-dev	32.6	74.3	42.0	35.0	74.8	44.8	29.0
Proposal-Free	AdaptIS [56]	val	35.9	-	-	40.3	-	-	29.3
	DeeperLab [63]	val	33.8	-	-	-	-	-	-
	DeeperLab [63]	test-dev	34.3	77.1	43.1	37.5	77.5	46.8	29.6
	SSAP [16]	val	36.5	-	-	-	-	-	-
	SSAP [16]	test-dev	36.9	80.7	44.8	40.1	81.6	48.5	32.0
	Ours (1x)	val	**37.51**	77.65	**47.18**	40.01	78.39	**50.03**	**33.74**
	Ours (1x)	test-dev	**37.7**	77.8	**47.3**	**40.7**	78.7	**50.7**	**33.1**

Qualitative results

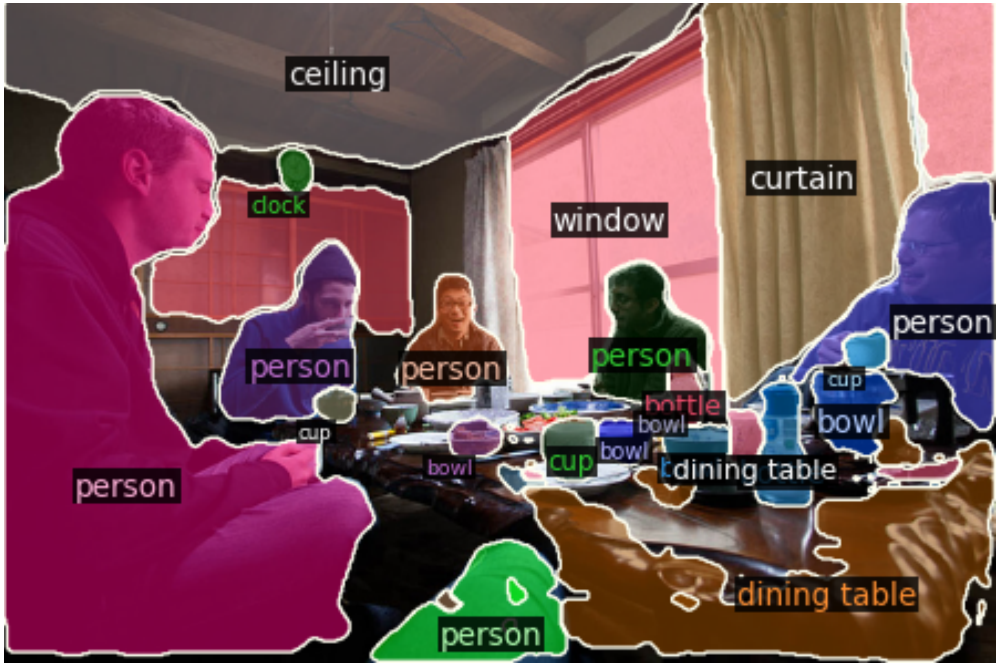



building
tree
person
teddy bear
handbag

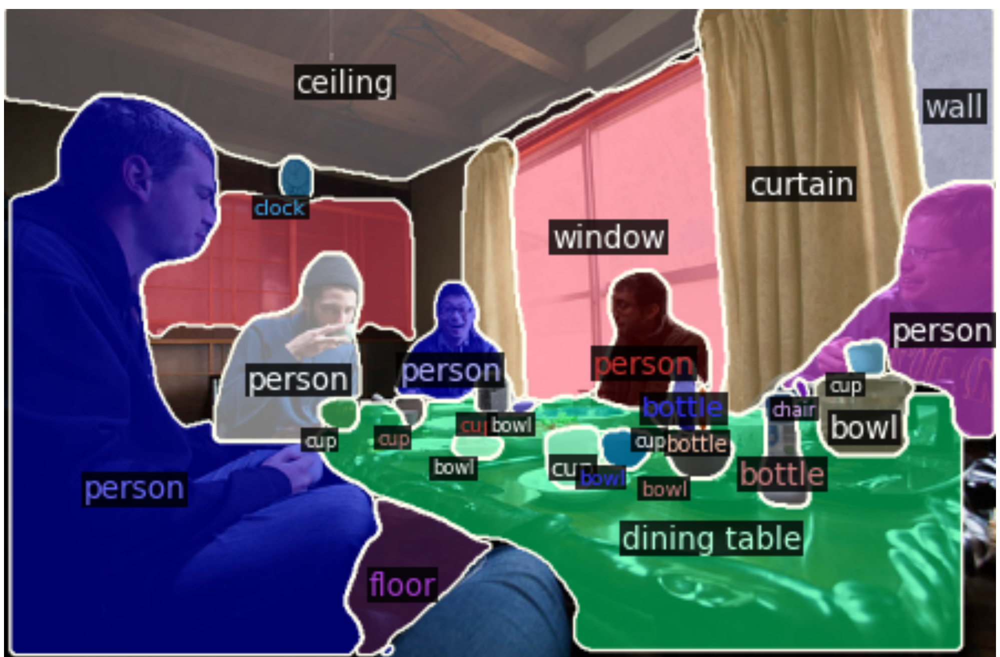
PCV


tree
person
teddy bear

UPSNet


ceiling
clock
curtain
window
person
person
person
person
person
cup
bottle
bowl
cup
bowl
bowl
cup
bowl
dining table
dining table
person
person

PCV


ceiling
wall
clock
curtain
window
person
person
person
person
person
cup
chair
bottle
bowl
cup
cup
bowl
cup
bottle
bowl
bottle
bowl
bowl
dining table
person
floor

UPSNet

Conclusion

- Instances emerge from pixel consensus on centroid locations
- Voting as dilated deconvolution
- Back-projection as convolutional filtering
- Training reduced to pixel labeling