



# Amodal and Panoptic Segmentation

Stephanie Liu, Andrew Zhou

# This lecture:

---

1. Semantic Amodal Segmentation
2. Cityscapes Dataset
3. ADE20K Dataset
4. Panoptic Segmentation

---

# Semantic Amodal Segmentation

Yan Zhu, Yuandong Tian, Dimitris Mexatas, and Piotr Dollar. “Semantic Amodal Segmentation” arXiv, 2016.

# Semantic Amodal Segmentation: Overview

---

- Motivation:
  - Train machines to see the “Invisible” (few has done so)
  - Amodal Annotation
  - Encourage researchers to use their dataset
- Central Questions:
  - Is amodal segmentation a well-posed annotation task?
  - Will multiple annotators agree on the annotation of a given image?
- YES.
  - Guidelines for annotators
  - Measures

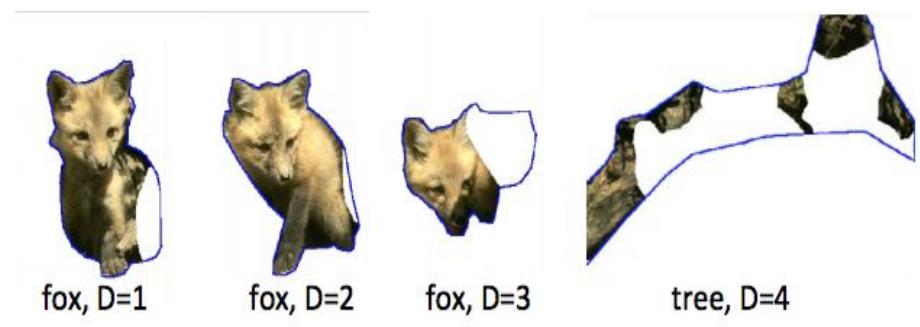


Image Credit: Yan Zhu et. al.

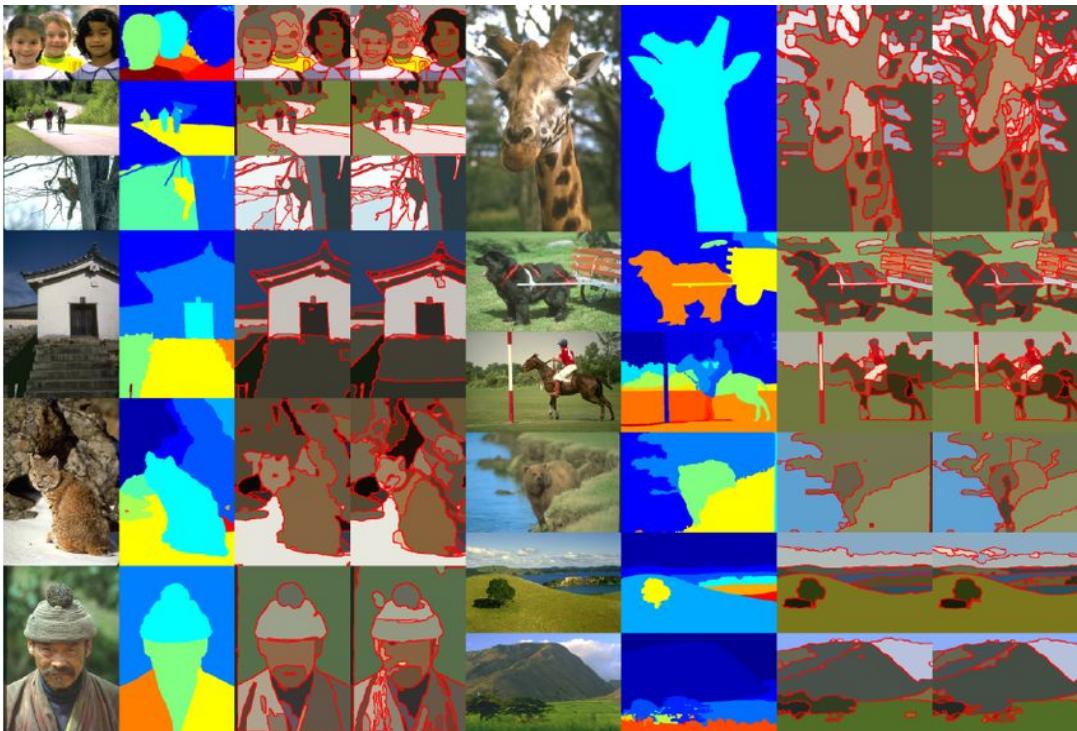


- Red: Modal Semantic Segmentation
- Green: Amodal Semantic Segmentation (Visible + Interpolated Regions)

# Datasets



# Berkeley Segmentation Dataset (BSDS)



- ❖ 1,000 Corel images
- ❖ 500 test images
- ❖ 12,000 hand-labeled segmentations
- ❖ Zhu et al. annotate 500 images

# MS COCO Dataset



- ❖ 328,000 images
- ❖ 2.5 million labeled instances
- ❖ Zhu et al. annotate 1,000 images

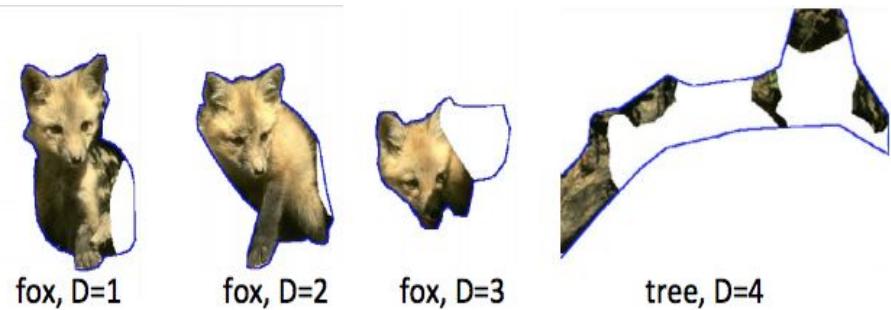
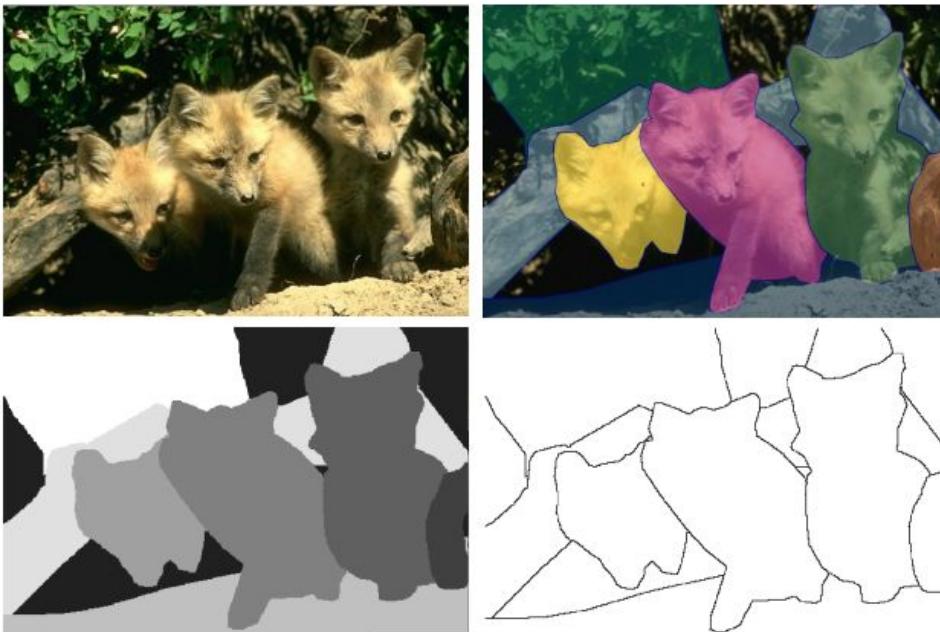
Image Credit: Tsung-Yi Lin et al

Slide Credit: Berthy Feng, Riley Simmons-Edler

# Four Guidelines For Annotation

---

# (1) Semantic Annotation



- Only annotate nameable regions

## (2) Dense Annotation

---

- All foreground object over a minimum size of 600 pixels should be labeled
- If an annotated region is occluded, occluder should also be annotated

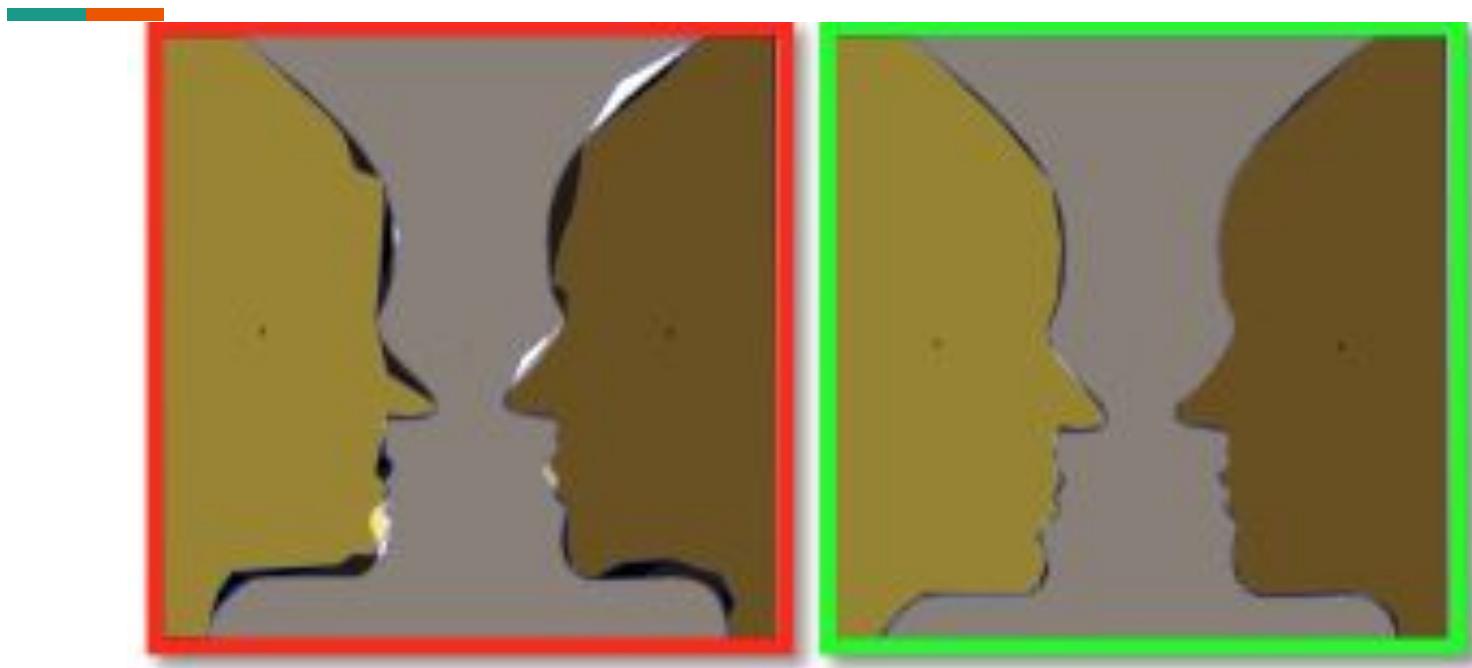
## (3) Depth Ordering

---



- Specify the relative depth ordering of all regions
- For non-overlapping regions any depth order is acceptable
- In ambiguous cases, depth order is specified so that edges are correctly ‘rendered’ (e.g., eyes go in front of the face)

## (4) Edge Sharing



- In figure-ground relation, edge belongs to foreground object
- When two regions are adjacent, annotator needs to mark shared edges, thus avoiding duplicate edges

# Dataset Statistics

- \* Analysis primarily based on BSDS



	BSDS COCO	
ann/image	5-7	1
regions/ann	7.3	9.2
points/region	64	46
pixel coverage	84%	69%
occlusion rate	62%	61%
occ/region	21%	31%
time/polygon	68s	41s
time/region	2m	2m
time/ann	15m	18m

(a) dataset summary statistics



(b) most common semantic labels

# Shape complexity

---

$$convexity(S) = \frac{Area(S)}{Area(ConvexHull(S))} \quad (1)$$

$$simplicity(S) = \frac{\sqrt{4\pi * Area(S)}}{Perimeter(S)} \quad (2)$$

	original	BSDS		COCO	
		modal	amodal	modal	amodal
simplicity	.801	.718	.834	.746	.856
convexity	.664	.616	.643	.658	.685
density	1.80%	1.57%	1.97%	1.71%	2.10%

→ More efficient to label than modal regions?

# Occlusion and Scene Complexity

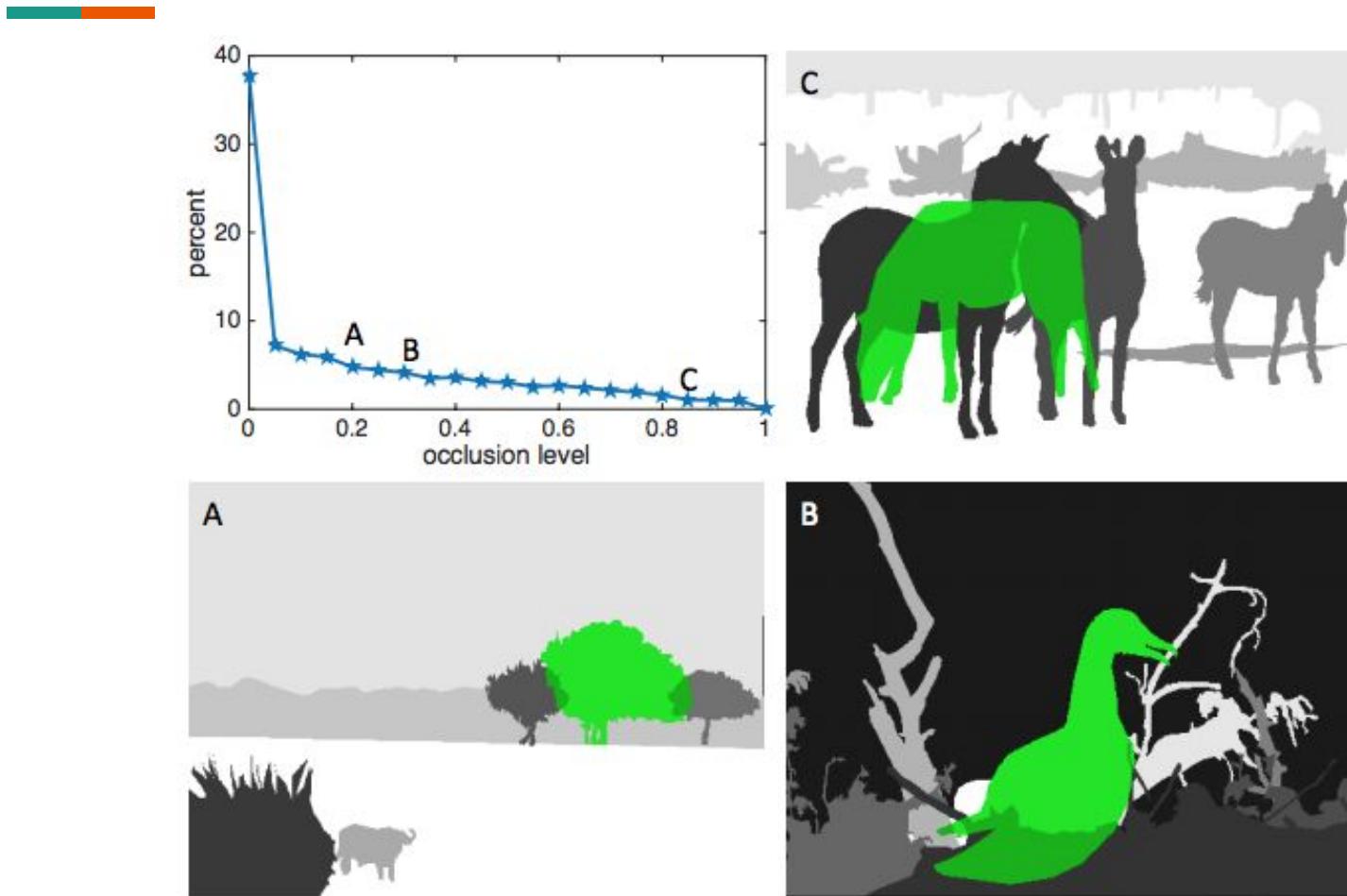
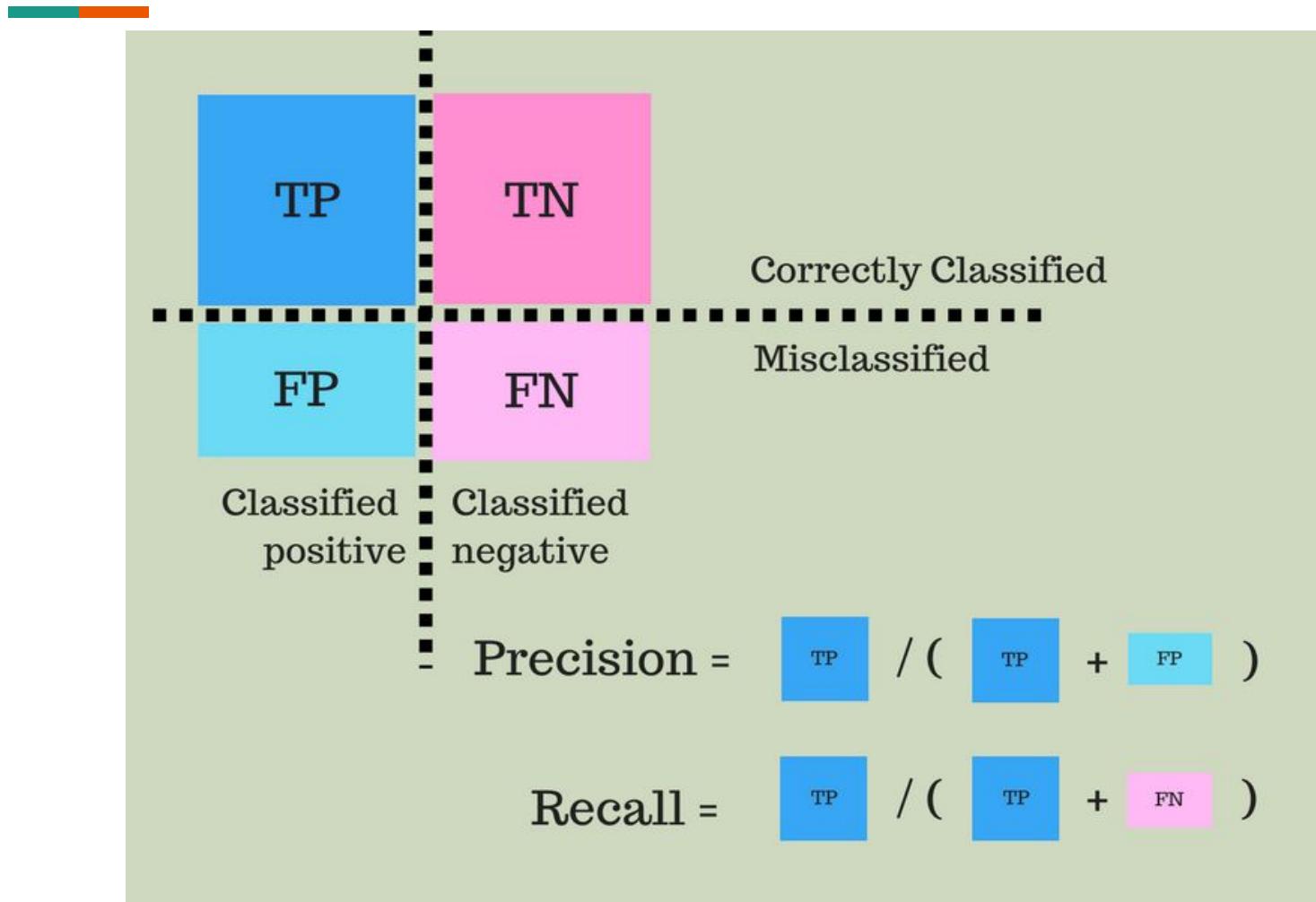


Image Credit: Yan Zhu et. al.

# Dataset Consistency

---

# Quick Review:



# Region Consistency

---

- $F = 2 P R / (P + R)$
- $n$  annotators yield  $n(n - 1)$  scores per image
- Paper amodal median: 0.723
- Original modal median: 0.425
- Paper modal median: 0.756



Image Credit: Yan Zhu et. al.

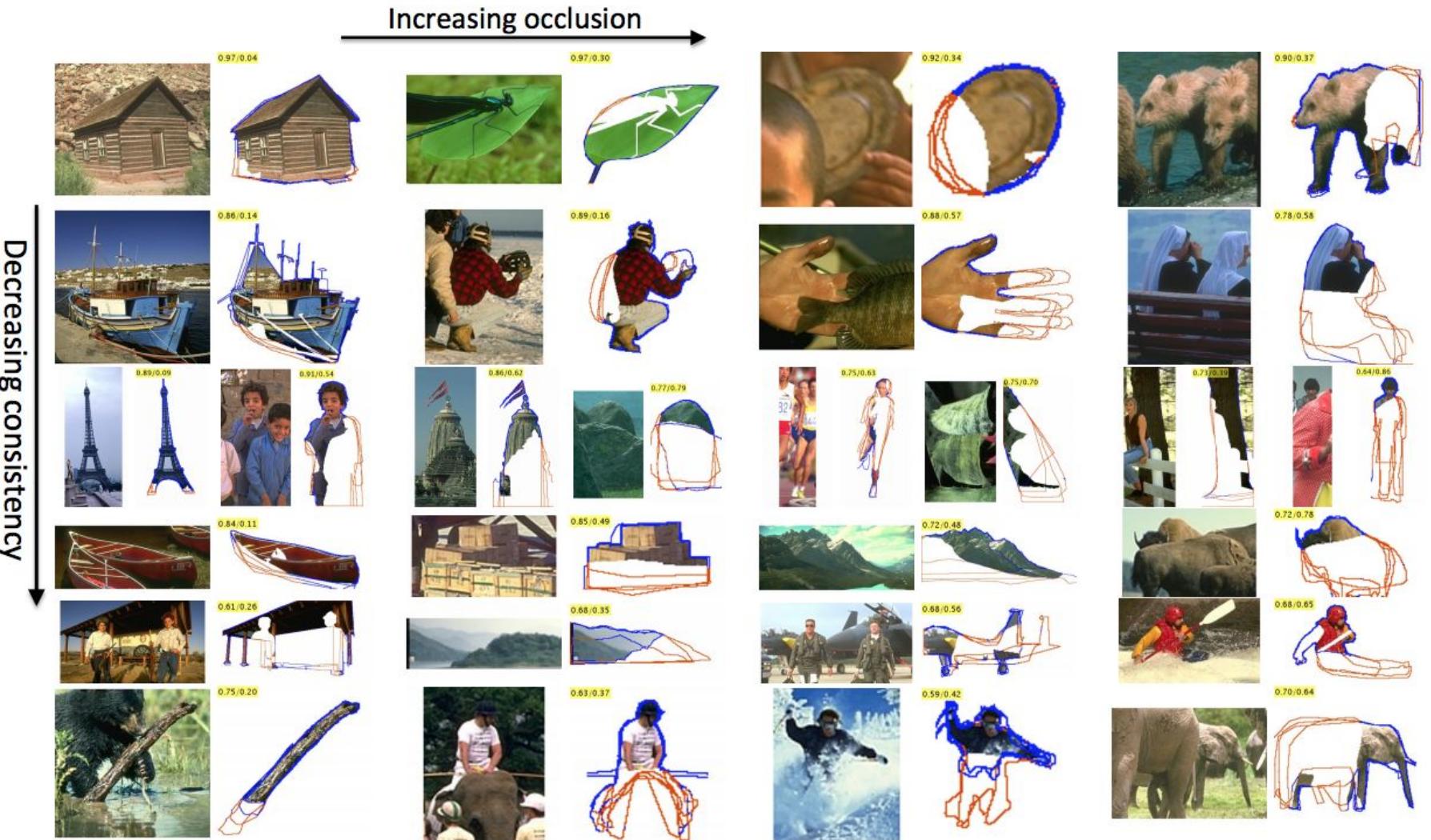


Image Credit: Yan Zhu et. al.

# Metrics and Baselines

---

# Amodal Segment Quality - Metrics

---

- Adopt *Average Recall* (AR) from COCO challenges
  - AR: segment recalls are computed at multiple IoU thresholds (0.5-0.95), then averaged
- Measure AR for 1000 segments per image
- Report AR for varying occlusion levels
  - N: none ( $q = 0$ )
  - P: partial ( $0 < q \leq 0.25$ )
  - H: heavy ( $0.25 < q$ )

# Amodal Segment Quality - Baselines

---

- *DeepMask* and *SharpMask*
- **ExpandMask:**
  - Input: image patch and modal mask generated by *SharpMask*
  - Output: amodal mask
- **AmodalMask:**
  - Directly predict amodal masks from image patches

# Amodal Segment Quality - Results

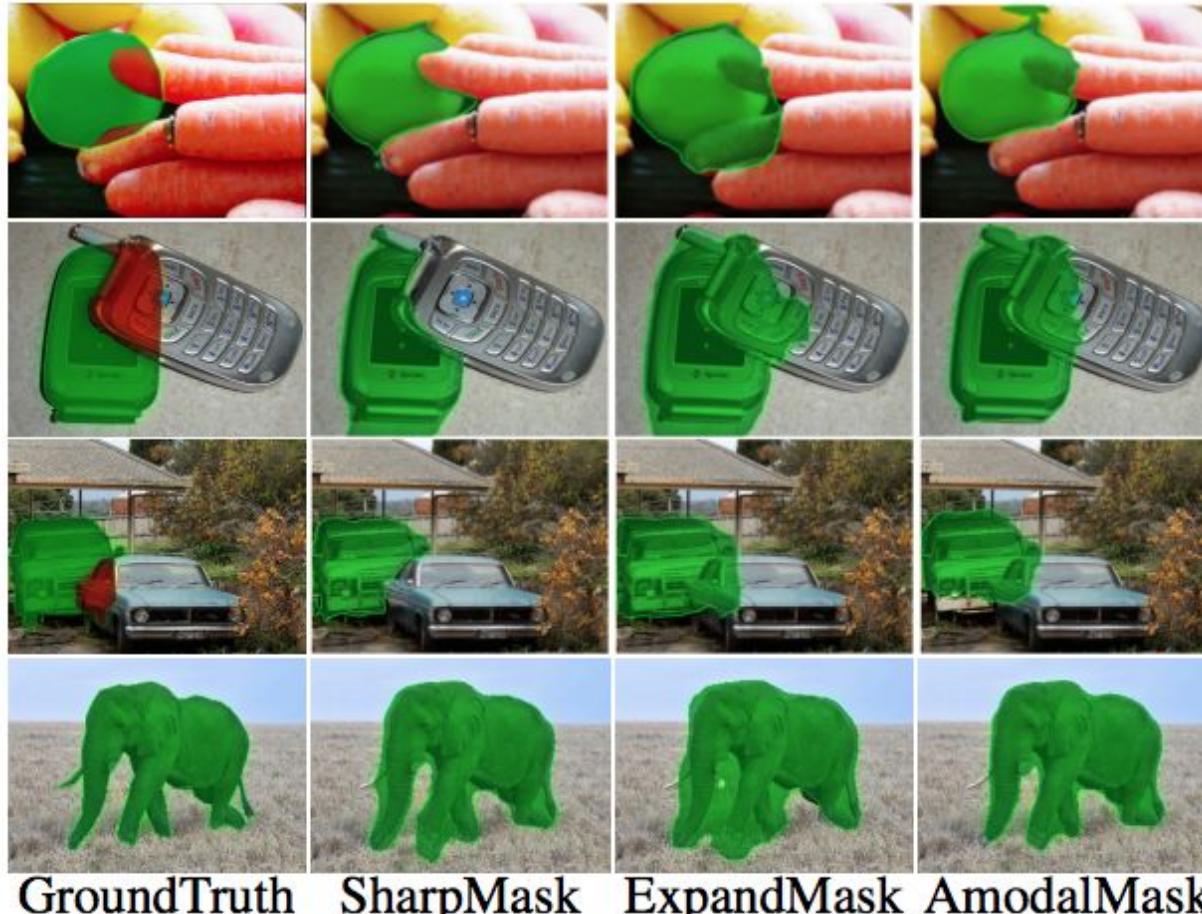


Image Credit: Yan Zhu et. al.

# Amodal Segment Quality - Results



	all regions				things only				stuff only			
	AR	AR <sup>N</sup>	AR <sup>P</sup>	AR <sup>H</sup>	AR	AR <sup>N</sup>	AR <sup>P</sup>	AR <sup>H</sup>	AR	AR <sup>N</sup>	AR <sup>P</sup>	AR <sup>H</sup>
DeepMask [31]	.378	.456	.407	.248	.422	.470	.473	.279	.248	.367	.242	.199
SharpMask [32]	.396	<b>.493</b>	.428	.242	.448	<b>.510</b>	<b>.501</b>	.275	.246	.384	.243	.187
ExpandMask <sup>S</sup>	.384	.460	.415	.256	.427	.474	.480	.284	.258	.374	.250	.212
AmodalMask <sup>S</sup>	.395	.457	.424	.289	.435	.468	.487	.316	.282	.388	.268	.246
ExpandMask	.417	.480	.428	.327	.456	.495	.488	.351	.305	.387	.278	.289
AmodalMask	<b>.434</b>	.470	<b>.460</b>	<b>.364</b>	<b>.458</b>	.479	.498	<b>.376</b>	<b>.366</b>	<b>.414</b>	<b>.365</b>	<b>.346</b>

---

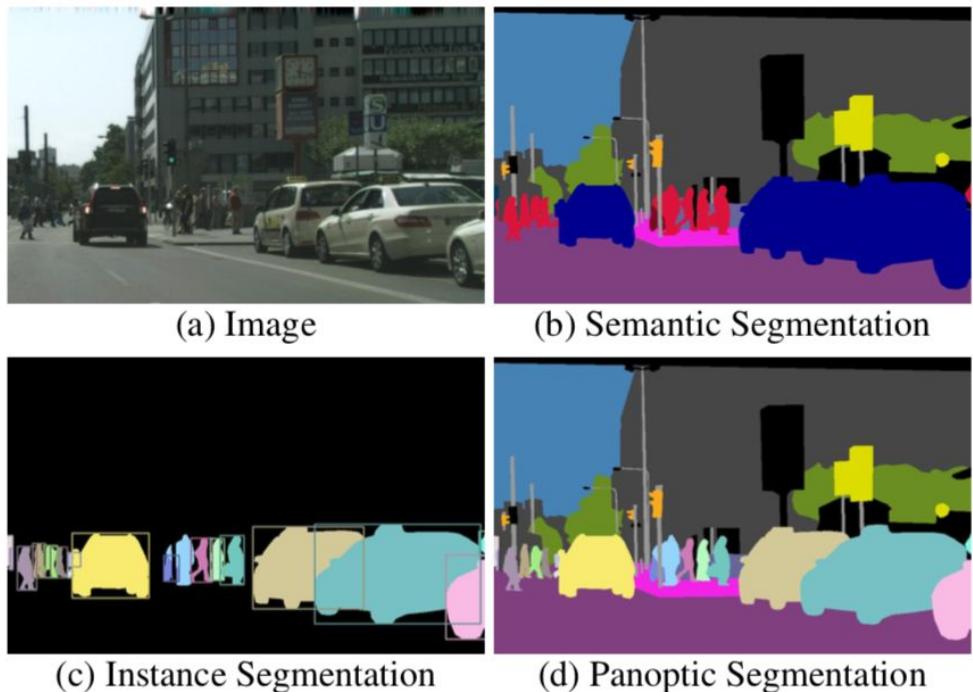
# The Cityscapes Dataset for Semantic Urban Scene Understanding

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld,  
Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and  
Bernt Schiele.

# Stuff v. Things

---

- **Stuff** = semantic segmentation
  - Assigning category label to each pixel
  - Grass, sky, road
- 
- **Things** = object detection or instance segmentation
  - Detect each object and delineate it
  - Car, person, chair



# Quick Review: Previous Datasets

- **PASCAL VOC**: bounding boxes around object and 20 classes
- **COCO**: focuses on instance segmentation
  - 2017 Stuff Segmentation Challenge (91 classes)

**Table 1.** Comparison of semantic segmentation datasets.

	Images	Obj. Inst.	Obj. Cls.	Part Inst.	Part Cls.	Obj. Cls. per Img.
COCO	123,287	886,284	91	0	0	3.5
ImageNet*	476,688	534,309	200	0	0	1.7
NYU Depth V2	1,449	34,064	894	0	0	14.1
→ Cityscapes	25,000	65,385	30	0	0	12.2
SUN	16,873	313,884	4,479	0	0	9.8
OpenSurfaces	22,214	71,460	160	0	0	N/A
PascalContext	10,103	~104,398**	540	181,770	40	5.1
→ ADE20K	22,210	434,826	2,693	175,961	476	9.9

\* has only bounding boxes (no pixel-level segmentation). Sparse annotations.

\*\* PascalContext dataset does not have instance segmentation. In order to estimate the number of instances, we find connected components (having at least 150pixels) for each class label.



# CITYSCAPES DATASET

---



- Both stuff and thing annotations
  - Captures the complexity of real-world urban scenes
- 
- 5,000 images with fine annotations
  - 20,000 with coarse annotations

# Collection & Evaluation

---

# Annotation

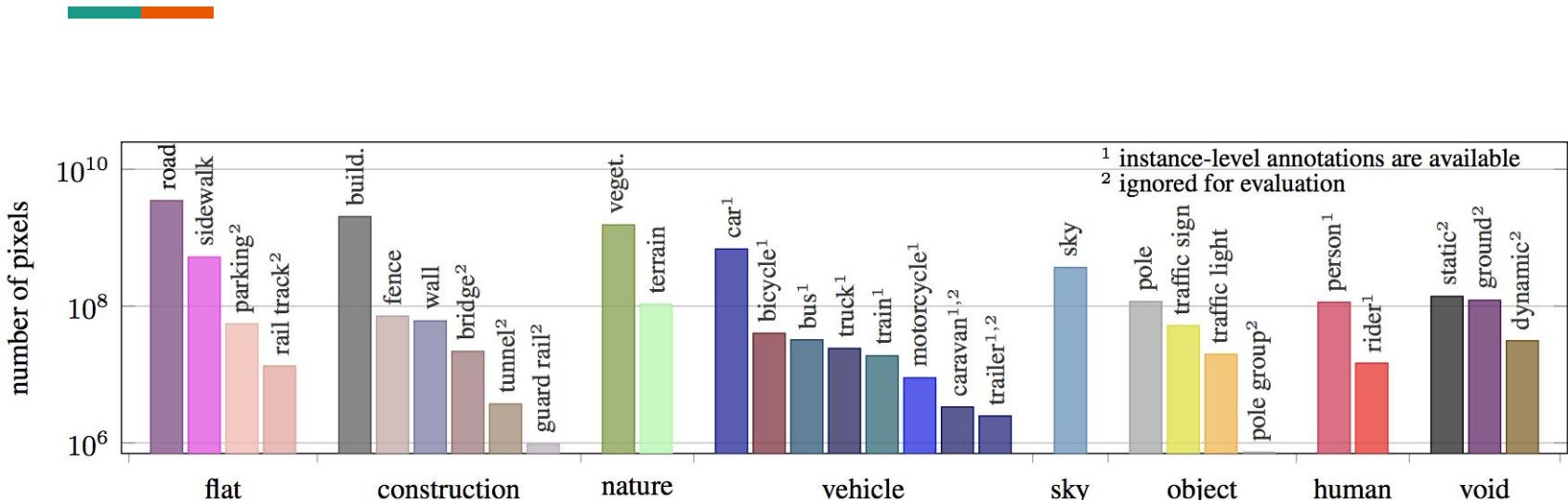


Figure 1. Number of finely annotated pixels (y-axis) per class and their associated categories (x-axis).

- Pixel-level and instance-level semantic labeling
  - Pixel-level: FCN model
  - Instance-level: FRCN to score object proposals
- 30 object classes, grouped into 8 categories

# Dataset Overview

- Goals
  - Annotation volume and density
  - Distribution of visual classes
  - Scene complexity
- 5000 images: 2975 train, 500 val, 1525 test

	#pixels [10 <sup>9</sup> ]	annot. density [%]
Ours (fine)	9.43	<b>97.1</b>
Ours (coarse)	<b>26.0</b>	67.5
CamVid	0.62	96.2
DUS	0.14	63.0
KITTI	0.23	88.9

Table 1. Absolute number and density of annotated pixels for Cityscapes, DUS, KITTI, and CamVid (upscaled to  $1280 \times 720$  pixels to maintain the original aspect ratio).

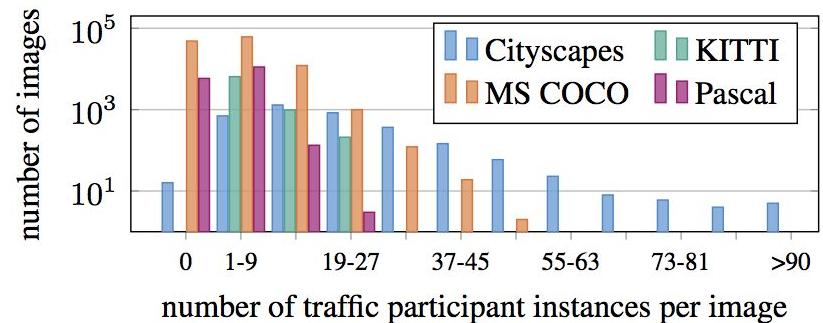


Figure 3. Dataset statistics regarding scene complexity. Only MS COCO and Cityscapes provide instance segmentation masks.

# Evaluation

---

- Cross-dataset evaluation
- Pixel-level semantic segmentation
  - Cityscapes: best-performing obtains IoU score of 67.1%
  - PASCAL VOC: 77.9%
- Instance-level semantic segmentation
  - Particularly challenging, AP score of 4.6%

Dataset	Best reported result	Our result
Camvid [7]	62.9 [4]	72.6
KITTI [58]	61.6 [4]	70.9
KITTI [64]	82.2 [73]	81.2

Table 5. Quantitative results (avg. recall in percent) of our half-resolution FCN-8s model trained on Cityscapes images and tested on Camvid and KITTI.

---

# Scene Parsing through ADE20K Dataset

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba

# ADE20K

---

- Focus on scene-parsing
  - 150 object and stuff classes
  - 20k training, 2k validation, 3k testing
- 
- Goal: collect a dataset that has pixel-level annotation with large open vocabulary



# Collection & Evaluation

---

# Annotation

- Dataset images
    - LabelMe
    - SUN
    - Places
  - Object and object parts
  - 82.4% consistency

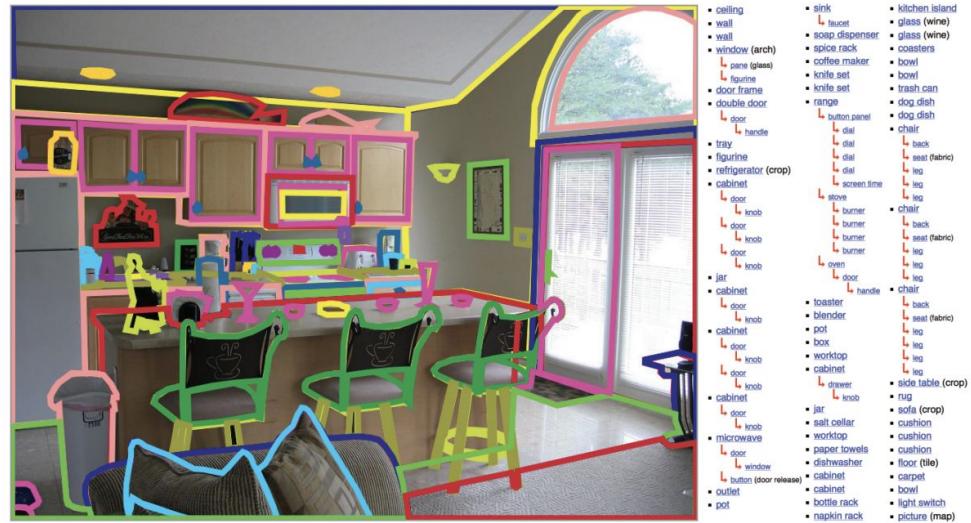
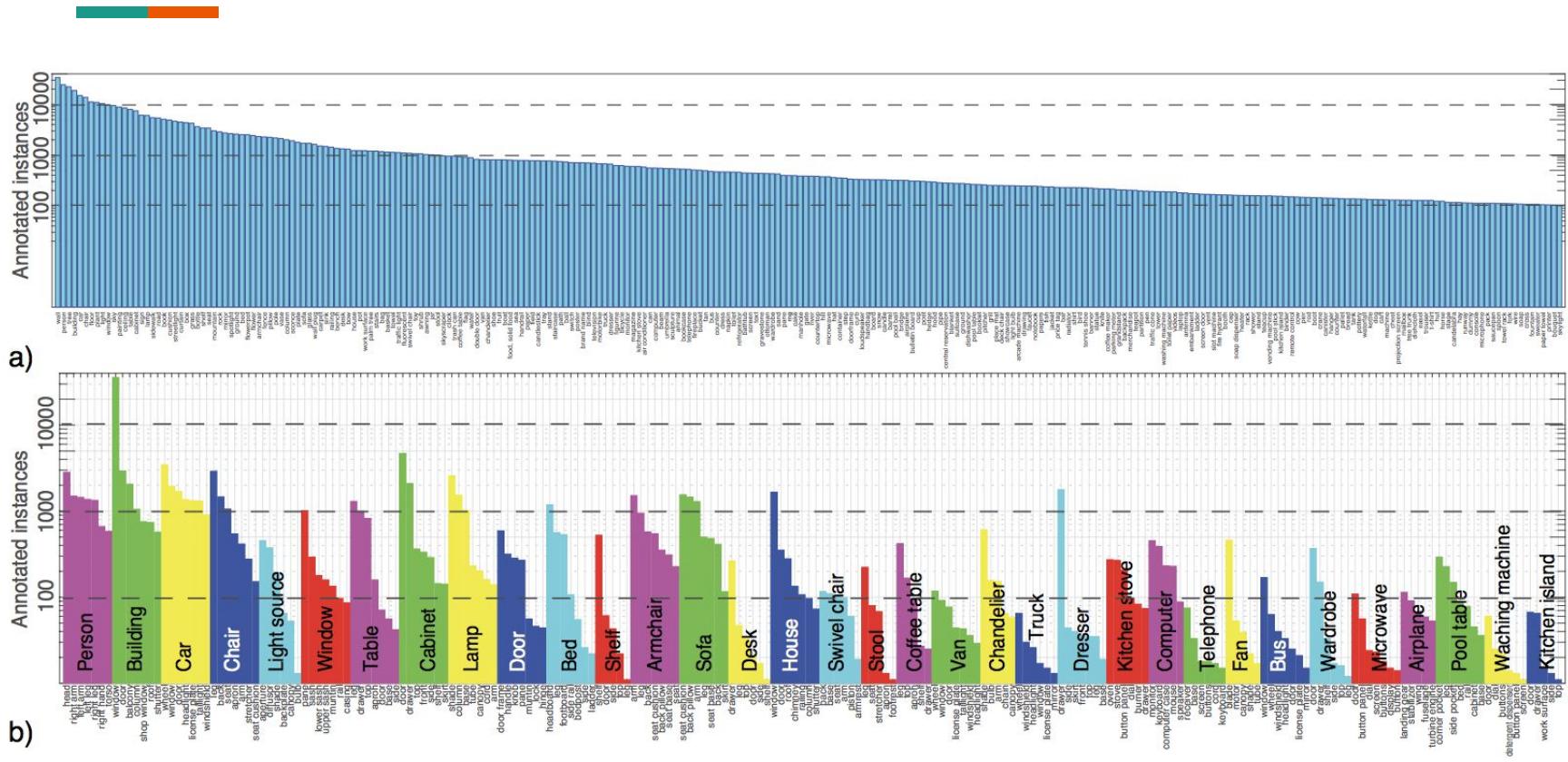
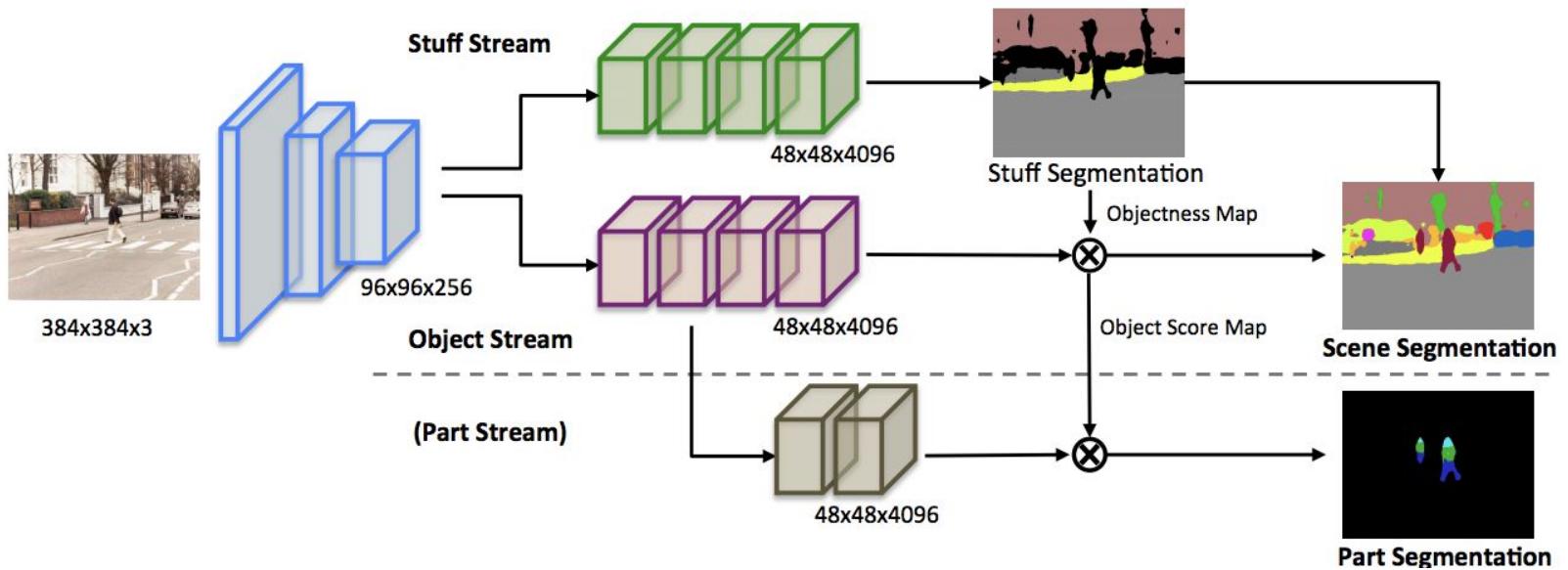


Figure 2. Annotation interface, the list of the objects and their associated parts in the image.

# Dataset Statistics



# Cascade Segmentation Module



# Evaluation

---

Table 1. Comparison of semantic segmentation datasets.

	Images	Obj. Inst.	Obj. Cls.	Part Inst.	Part Cls.	Obj. Cls. per Img.
COCO	123,287	886,284	91	0	0	3.5
ImageNet*	476,688	534,309	200	0	0	1.7
NYU Depth V2	1,449	34,064	894	0	0	14.1
Cityscapes	25,000	65,385	30	0	0	12.2
SUN	16,873	313,884	4,479	0	0	9.8
OpenSurfaces	22,214	71,460	160	0	0	N/A
PascalContext	10,103	~104,398**	540	181,770	40	5.1
ADE20K	22,210	434,826	2,693	175,961	476	9.9

\* has only bounding boxes (no pixel-level segmentation). Sparse annotations.

\*\* PascalContext dataset does not have instance segmentation. In order to estimate the number of instances, we find connected components (having at least 150pixels) for each class label.

- Compared to COCO and ImageNet, much more diverse scenes
- High annotation complexity

---

# Panoptic Segmentation

Kirillov, Alexander, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár

# Unifying Semantic and Instance Segmentation



# Unifying Semantic and Instance Segmentation



Semantic Segmentation

# Unifying Semantic and Instance Segmentation



Semantic Segmentation

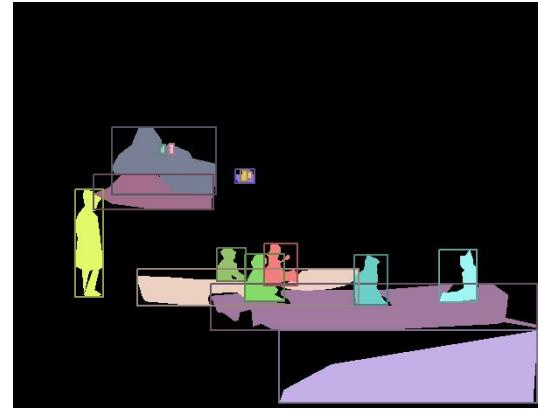


Object Detection

# Unifying Semantic and Instance Segmentation



Semantic Segmentation



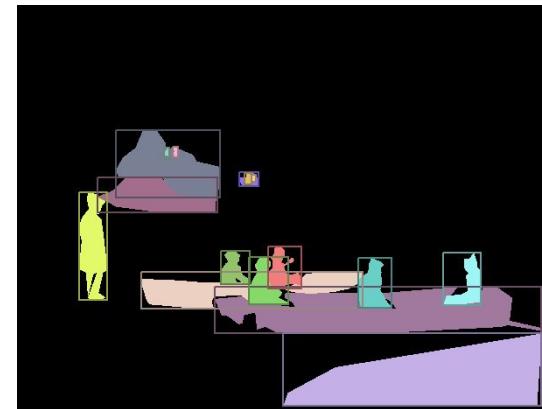
Object Detection/Seg

# Unifying Semantic and Instance Segmentation



**Semantic Segmentation**

- per-pixel annotation
- simple accuracy measure
- instances indistinguishable



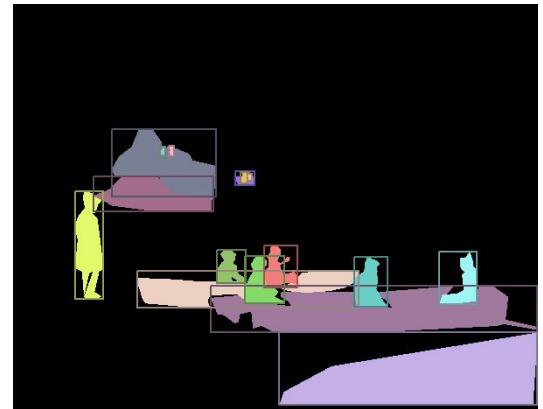
**Object Detection/Seg**

# Unifying Semantic and Instance Segmentation



**Semantic Segmentation**

- per-pixel annotation
- simple accuracy measure
- instances indistinguishable



**Object Detection/Seg**

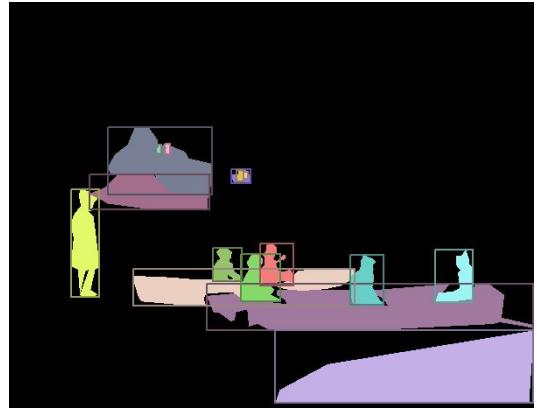
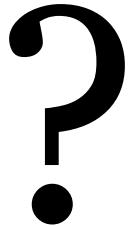
- each object detected and segmented separately
- “stuff” is not segmented

# Unifying Semantic and Instance Segmentation



**Semantic Segmentation**

- per-pixel annotation
- simple accuracy measure
- instances indistinguishable



**Object Detection/Seg**

- each object detected and segmented separately
- “stuff” is not segmented

# Unifying Semantic and Instance Segmentation

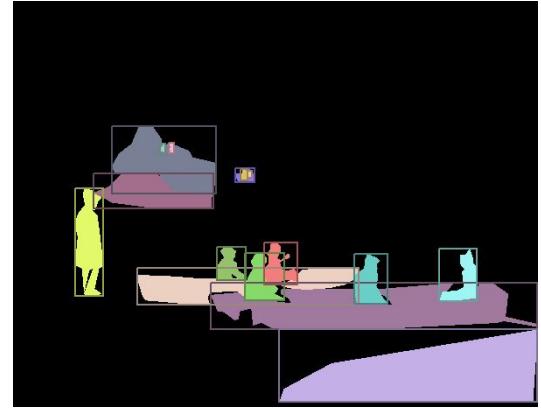


**Semantic Segmentation**

- per-pixel annotation
- simple accuracy measure
- instances indistinguishable



**Panoptic Segmentation**



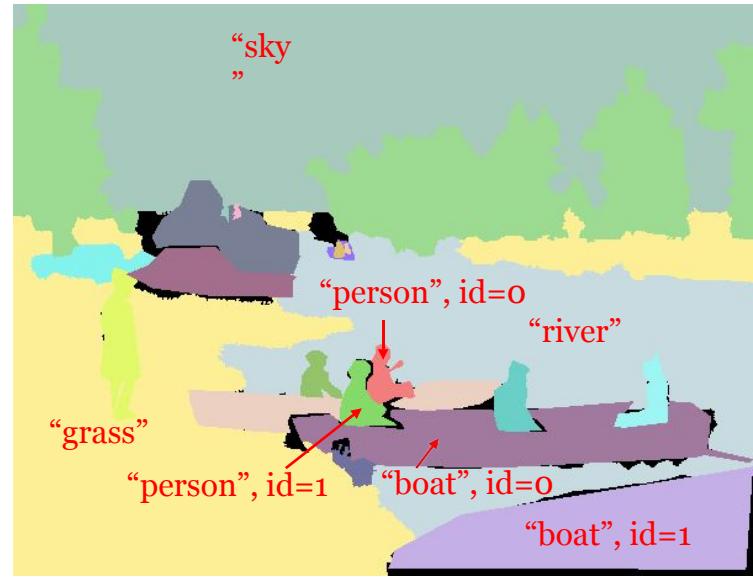
**Object Detection/Seg**

- each object detected and segmented separately
- “stuff” is not segmented

# Outline

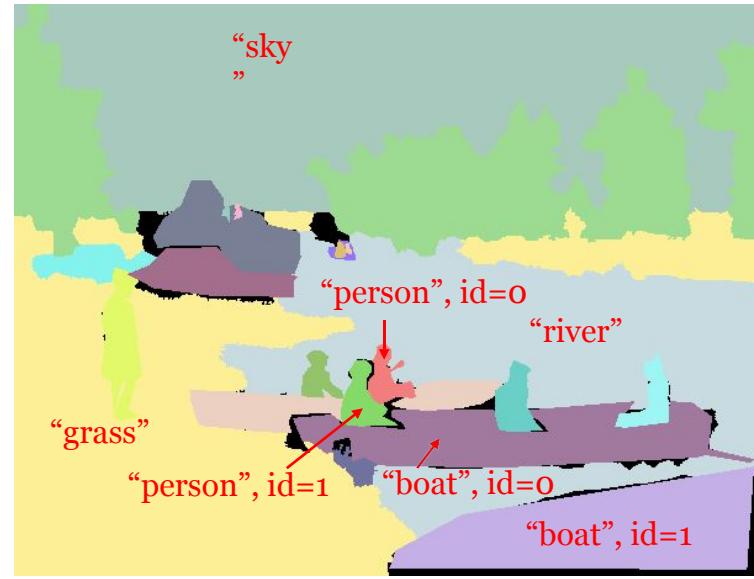
- Motivation
- **Problem Definition**
- Quality Evaluation
- Human Performance
- Humans vs Computers
- Perspectives

# Panoptic Segmentation



For each pixel  $i$  predict semantic label  $l$  and instance id  $z$

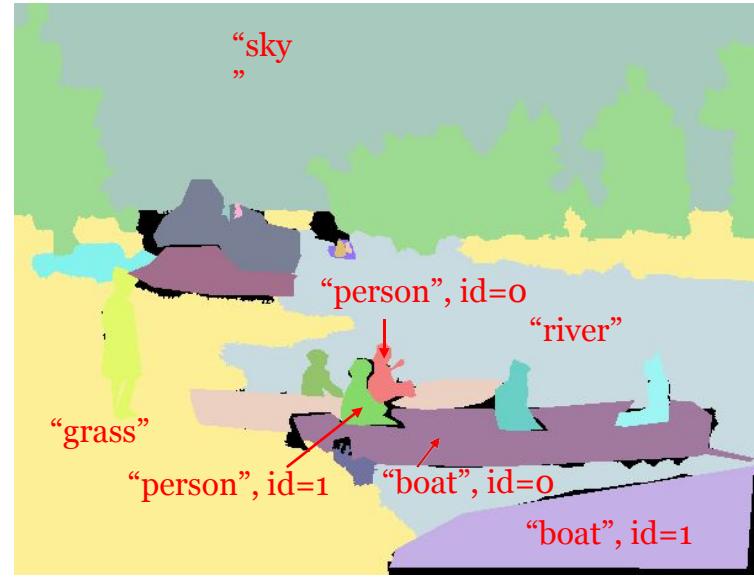
# Panoptic Segmentation



For each pixel  $i$  predict semantic label  $l$  and instance id  $z$

➤ no overlaps between segments

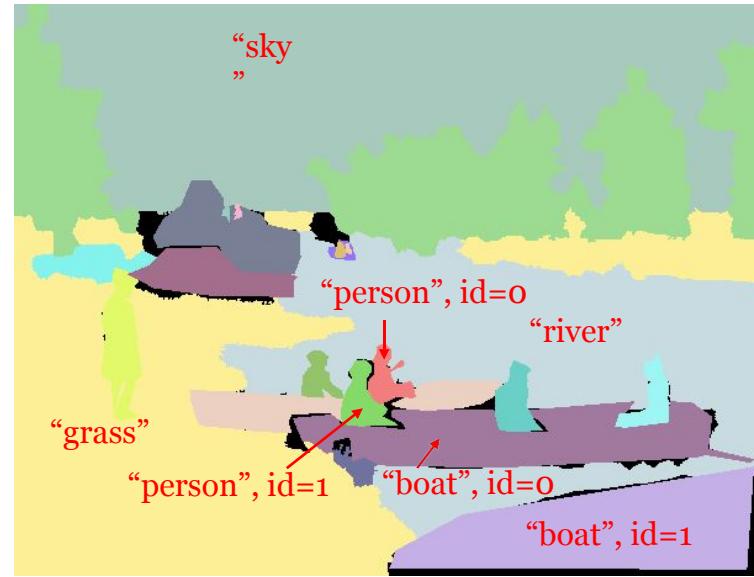
# Panoptic Segmentation



For each pixel  $i$  predict semantic label  $l$  and instance id  $z$   
➤ no overlaps between segments

- Popular datasets can be used
- Introduce simple, intuitive metric
- Drive novel algorithmic ideas

# Popular datasets can be used



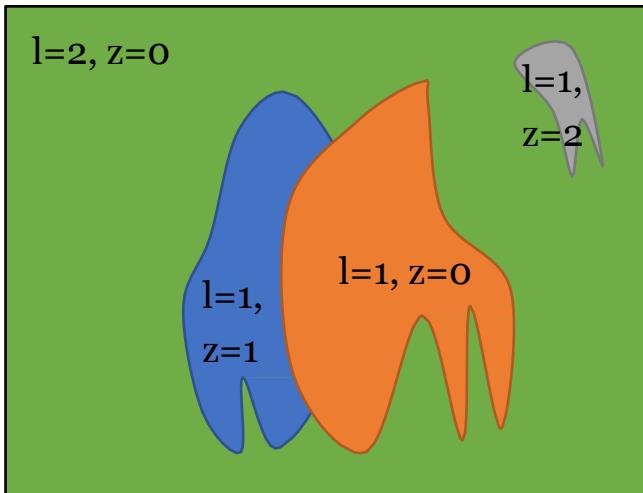
For each pixel  $i$  predict semantic label  $l$  and instance id  $z$

Datasets	Instance Segmentation	Semantic Segmentation
COCO*	+	+
ADE20k/Places	+	+
CityScapes	+	+
Mapillary Vistas	+	+

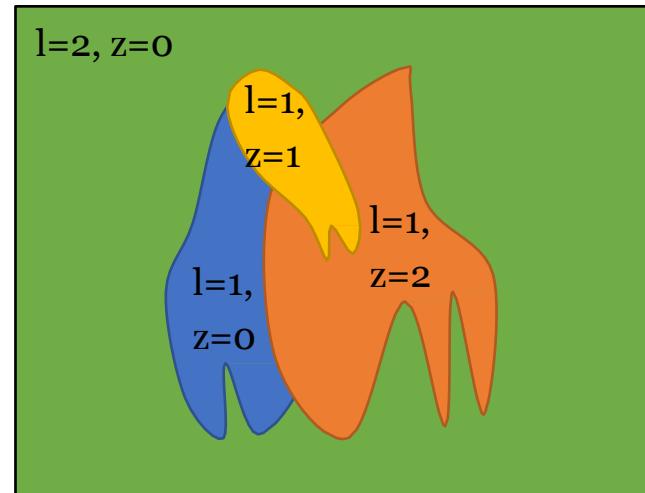
# Outline

- Motivation
- Problem Definition
- **Quality Evaluation**
- Human Performance
- Humans vs Computers
- Perspectives

# Quality Evaluation

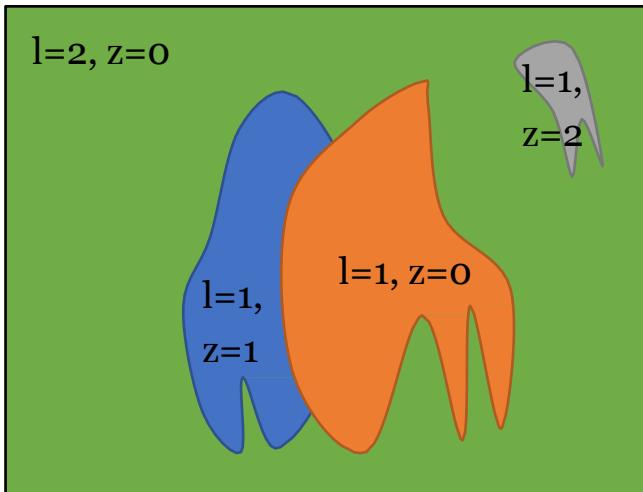


Ground Truth

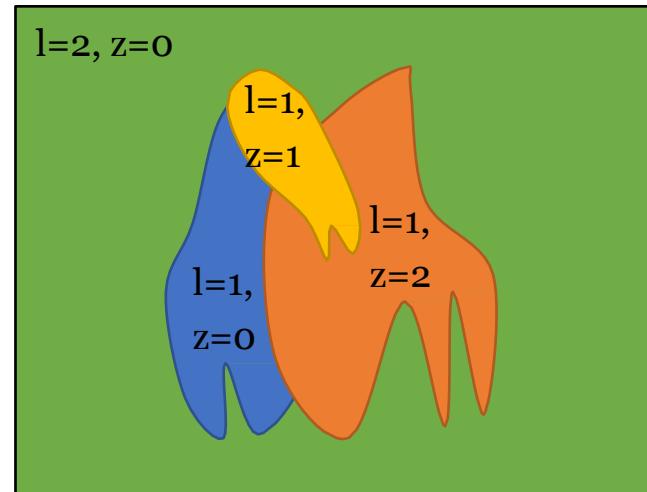


Prediction

# Quality Evaluation



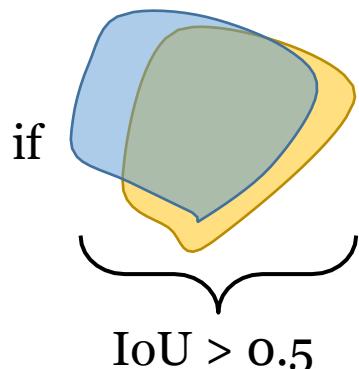
Ground Truth



Prediction

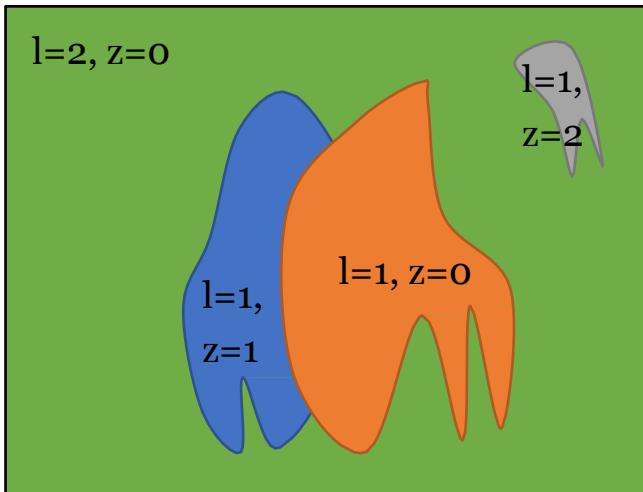
**Theorem:** Matching is unique if overlapping threshold  $> 0.5$  IoU and both ground truth and prediction have no overlaps.

**Proof sketch:**

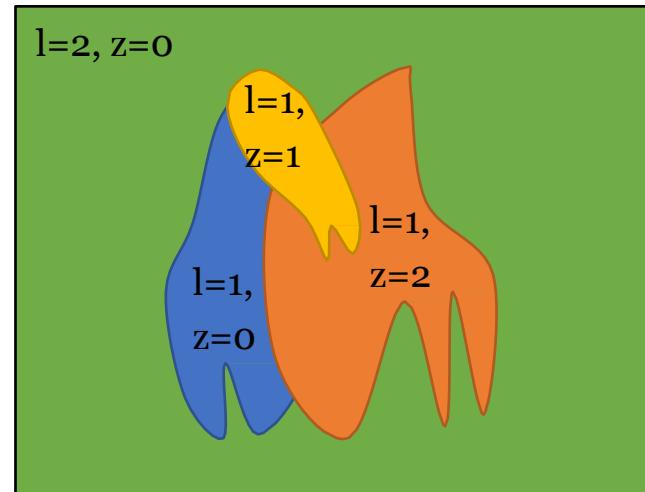


if  
then there is no other non overlapping object that has  $\text{IoU} > 0.5$ .

# Quality Evaluation



Ground Truth



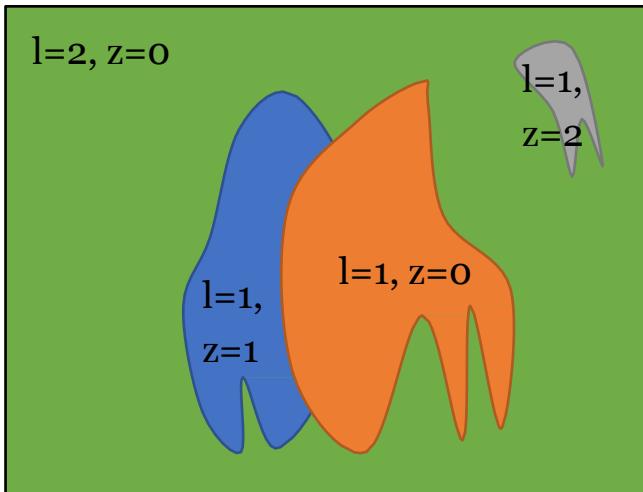
Prediction

$$TP_1 = \{(\boxed{\text{blue blob}}, \boxed{\text{blue blob}}), (\boxed{\text{orange flame}}, \boxed{\text{orange flame}})\}$$

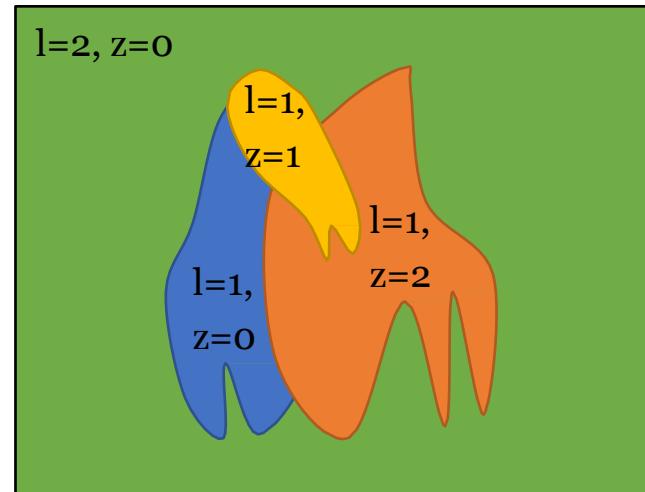
$$FP_1 = \{\boxed{\text{yellow blob}}\}$$

$$FN_1 = \{\boxed{\text{grey blob}}\}$$

# Quality Evaluation



Ground Truth



Prediction

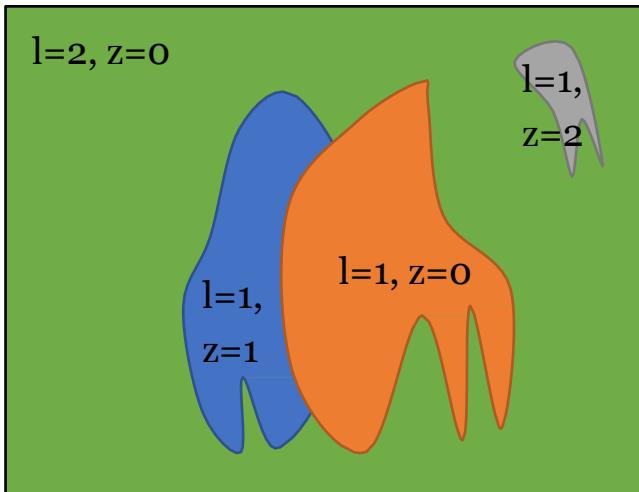
$$TP_1 = \{(\boxed{\text{blue}}), \boxed{\text{blue}}), (\boxed{\text{orange}}), \boxed{\text{orange}})\}$$

$$FP_1 = \{\boxed{\text{yellow}}\}$$

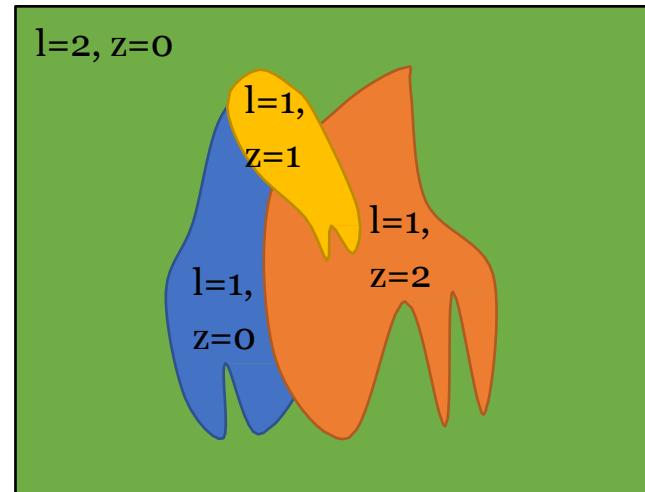
$$FN_1 = \{\boxed{\text{grey}}\}$$

$$PSQ_1 = \frac{IoU(\boxed{\text{blue}}, \boxed{\text{blue}}) + IoU(\boxed{\text{orange}}, \boxed{\text{orange}})}{|TP_1| + |FP_1| + |FN_1|} = \frac{\sum_{(g,p) \in TP_1} IoU(g,p)}{|TP_1| + |FP_1| + |FN_1|}$$

# Quality Evaluation



Ground Truth

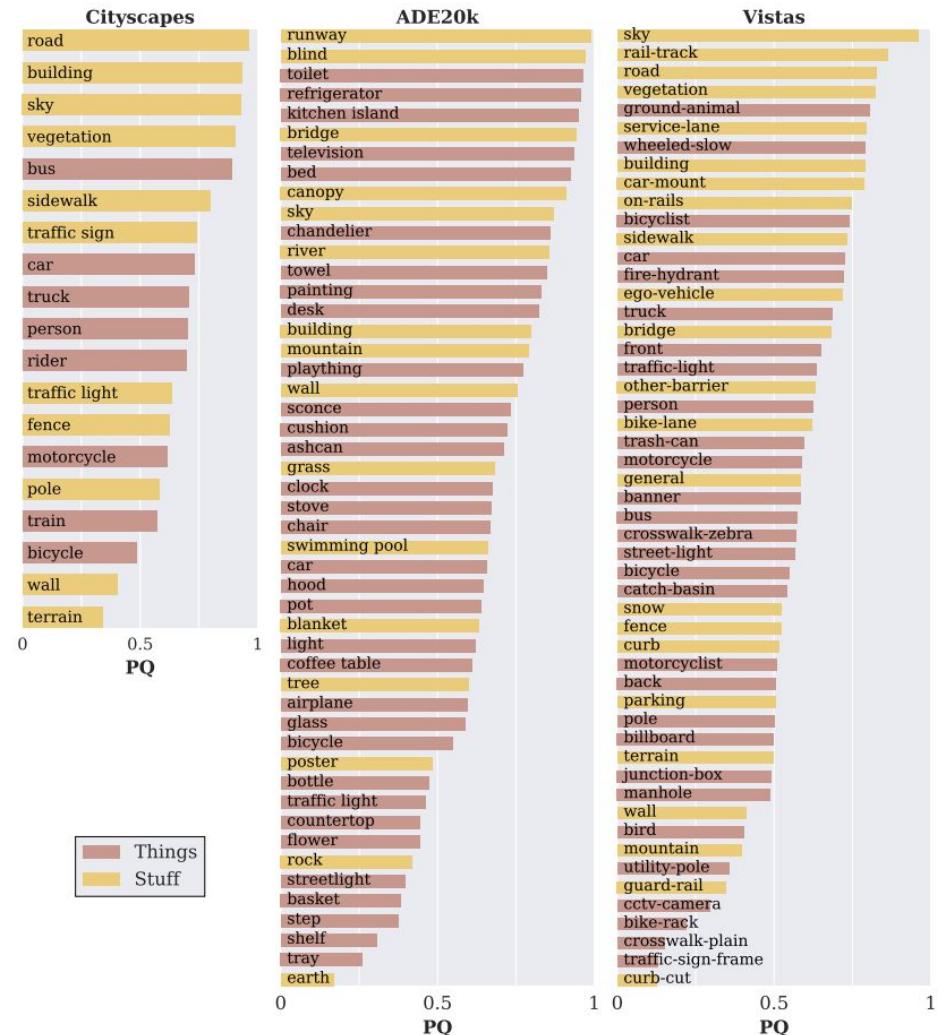


Predictio  
n

$$\text{PSQ}_l = \frac{\sum_{(g,p) \in \text{TP}_l} \text{IoU}(g,p)}{|\text{TP}_l| + |\text{FP}_l| + |\text{FN}_l|} = \underbrace{\frac{\sum_{(g,p) \in \text{TP}_l} \text{IoU}(g,p)}{|\text{TP}_l|}}_{\text{Segmentation Quality}} \cdot \underbrace{\frac{|\text{TP}_l|}{|\text{TP}_l| + |\text{FP}_l| + |\text{FN}_l|}}_{\text{Detection Quality}}$$

# Quality Evaluation

Things and stuff are distributed evenly  
 => PQ balances performance



**Figure 5: Per-Class Human performance, sorted by PQ.** Thing classes are shown in red, stuff classes in orange (for ADE20k every other class is shown, classes without matches in the dual-annotated test sets are omitted). Things and stuff are distributed fairly evenly, implying PQ balances their performance.

# Outline

- Motivation
- Problem Definition
- Quality Evaluation
- **Human Performance**
- Humans vs Computers
- Perspectives

# Panoptic Segmentation Quality (PSQ)

$$\text{PSQ}_l = \frac{\sum_{(g,p) \in \text{TP}_l} \text{IoU}(g,p)}{|\text{TP}_l| + |\text{FP}_l| + |\text{FN}_l|} = \underbrace{\frac{\sum_{(g,p) \in \text{TP}_l} \text{IoU}(g,p)}{|\text{TP}_l|}}_{\text{Seg Quality}} \cdot \underbrace{\frac{|\text{TP}_{l=1}|}{|\text{TP}_l| + |\text{FP}_l| + |\text{FN}_l|}}_{\text{Det Quality}}$$

no confidence scores  
↓  
human performance  
can be measured

# Panoptic Segmentation Quality (PSQ)

$$\text{PSQ}_l = \frac{\sum_{(g,p) \in \text{TP}_l} \text{IoU}(g,p)}{|\text{TP}_l| + |\text{FP}_l| + |\text{FN}_l|} = \underbrace{\frac{\sum_{(g,p) \in \text{TP}_l} \text{IoU}(g,p)}{|\text{TP}_l|}}_{\text{Seg Quality}} \cdot \underbrace{\frac{|\text{TP}_{l=1}|}{|\text{TP}_l| + |\text{FP}_l| + |\text{FN}_l|}}_{\text{Det Quality}}$$

no confidence scores  
↓  
human performance  
can be measured

**CityScapes:** 30 images were annotated independently twice.

# Panoptic Segmentation Quality (PSQ)

$$\text{PSQ}_1 = \frac{\sum_{(g,p) \in \text{TP}_1} \text{IoU}(g,p)}{|\text{TP}_1| + |\text{FP}_1| + |\text{FN}_1|} = \underbrace{\frac{\sum_{(g,p) \in \text{TP}_1} \text{IoU}(g,p)}{|\text{TP}_1|}}_{\text{Seg Quality}} \cdot \underbrace{\frac{|\text{TP}_{1=1}|}{|\text{TP}_1| + |\text{FP}_1| + |\text{FN}_1|}}_{\text{Det Quality}}$$

no confidence scores  
↓  
human performance  
can be measured

**CityScapes:** 30 images were annotated independently twice.

class	PSQ	Seg Quality	Det Quality
car	66.6%	87.5%	76.2%
person	61.8%	80.8%	76.4%
motorcycle	51.8%	77.8%	66.7%
pole	46.9%	70.3%	66.7%
road	98.0%	98.0%	100.0%
traffic sign	67.1%	79.5%	84.4%
average	62.6%	83.9%	73.43%

All Objects

# Panoptic Segmentation Quality (PSQ)

$$\text{PSQ}_1 = \frac{\sum_{(g,p) \in \text{TP}_1} \text{IoU}(g,p)}{|\text{TP}_1| + |\text{FP}_1| + |\text{FN}_1|} = \underbrace{\frac{\sum_{(g,p) \in \text{TP}_1} \text{IoU}(g,p)}{|\text{TP}_1|}}_{\text{Seg Quality}} \cdot \underbrace{\frac{|\text{TP}_{1=1}|}{|\text{TP}_1| + |\text{FP}_1| + |\text{FN}_1|}}_{\text{Det Quality}}$$

no confidence scores  
↓  
human performance  
can be measured

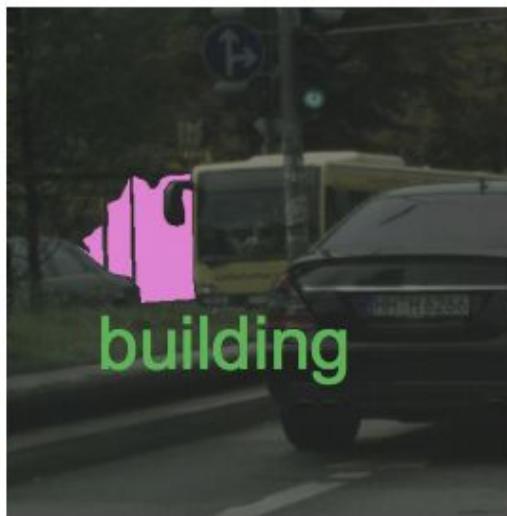
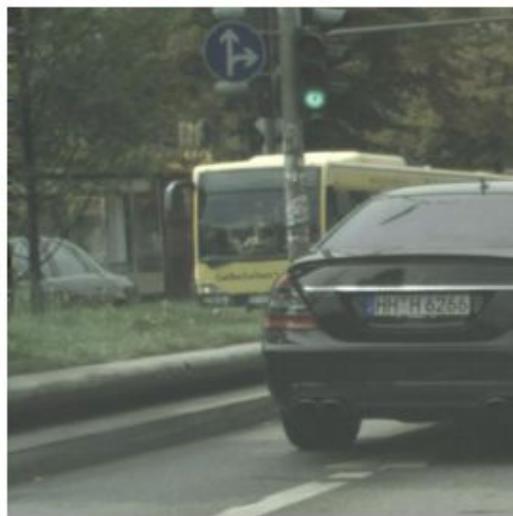
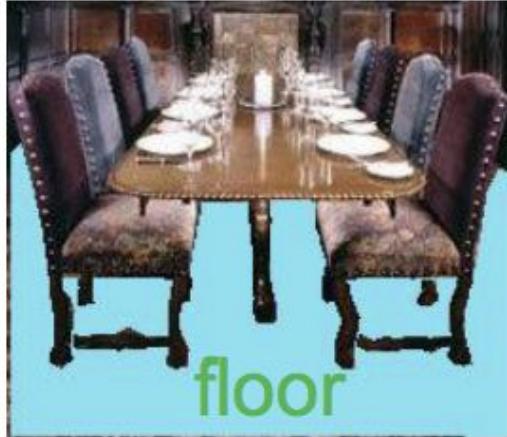
**CityScapes:** 30 images were annotated independently twice.

class	PSQ	Seg Quality	Det Quality
car	89.4%	91.3%	97.9%
person	82.0%	78.1%	94.1%
motorcycle	68.8%	79.4%	86.7%
pole	48.2%	70.3%	68.6%
road	98.0%	98.0%	100.0%
traffic sign	74.0%	79.5%	93.1%
average	68.7%	85.1%	80.1%

Objects >  $32^2$

# Human Annotation

Γ



## Classification Flaws

Image Credit: Alexander Kirillov et. al.

# Human Annotation

F

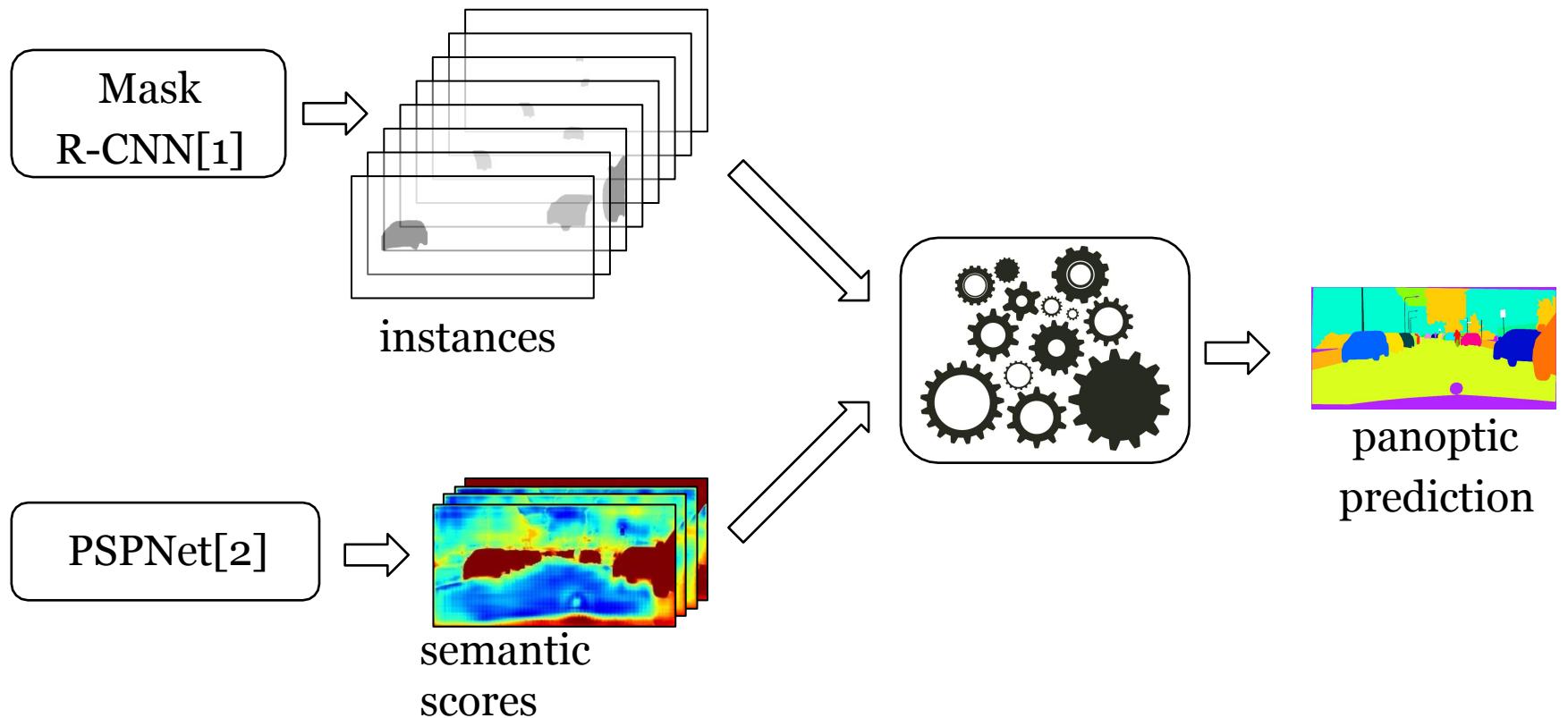


Segmentation Flaws

# Outline

- Motivation
- Problem Definition
- Quality Evaluation
- Human Performance
- **Humans vs Computers**
- Perspectives

# Mask R-CNN + PSPNet Combination Heuristic

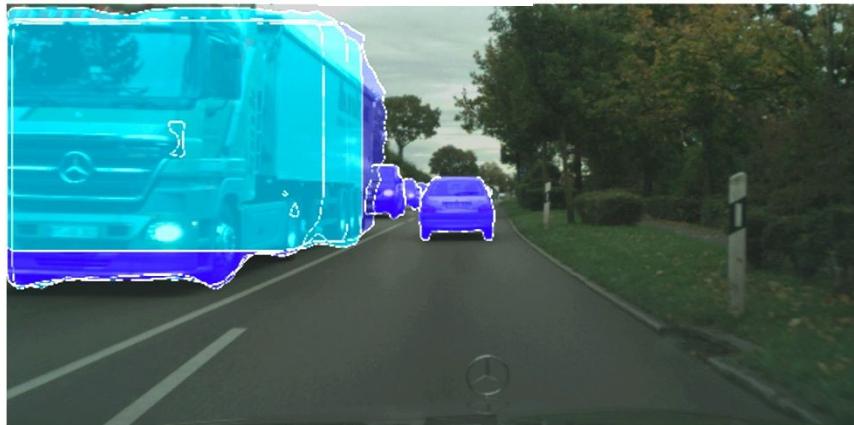


Slide Credit: Alexander Kirillov

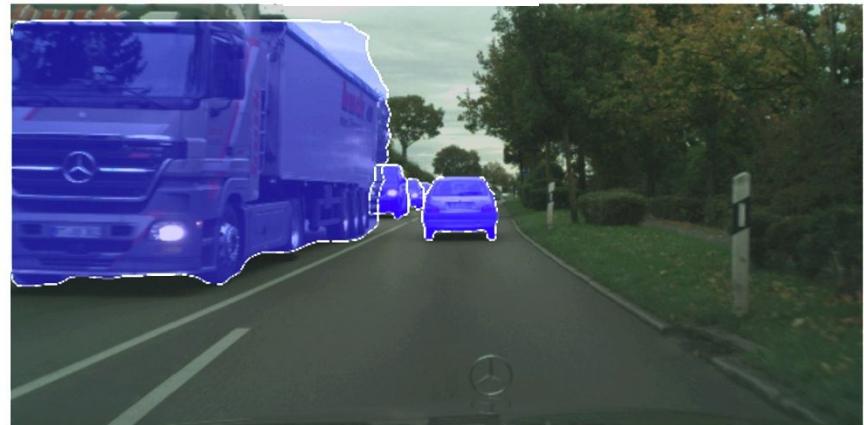
1 He, K., Gkioxari, G., Dollár, P., & Girshick, R. Mask R-CNN. ICCV 2017.

2 Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. Pyramid scene parsing network. CVPR 2017.

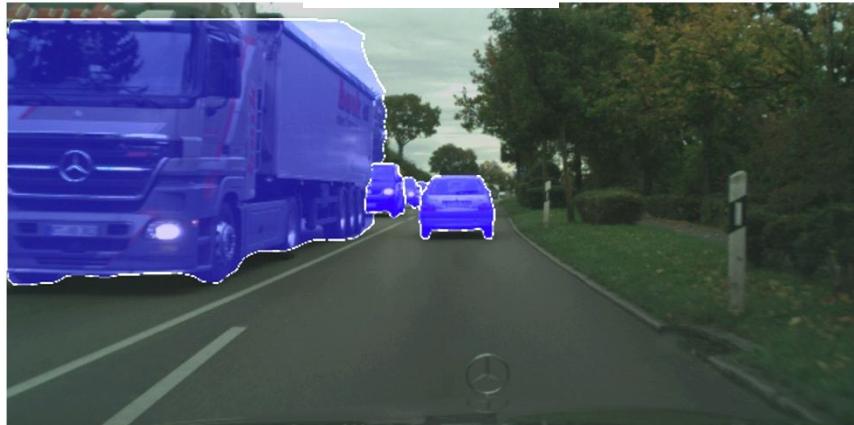
# Mask R-CNN Non-overlapping Instances



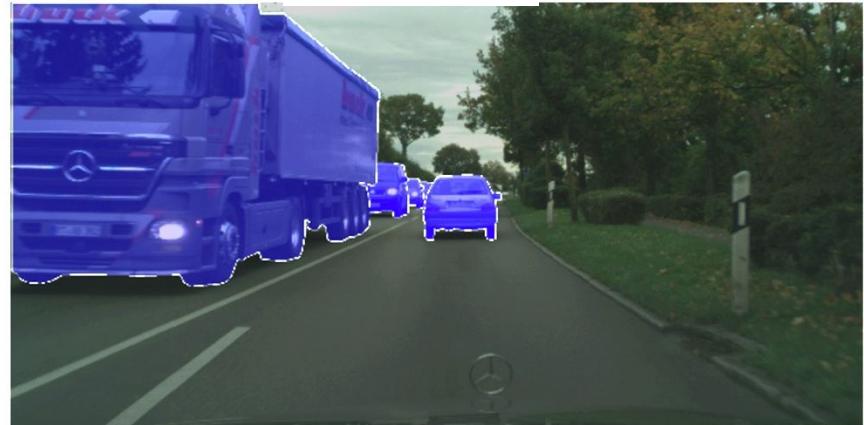
Mask R-CNN output



Mask R-CNN filtered



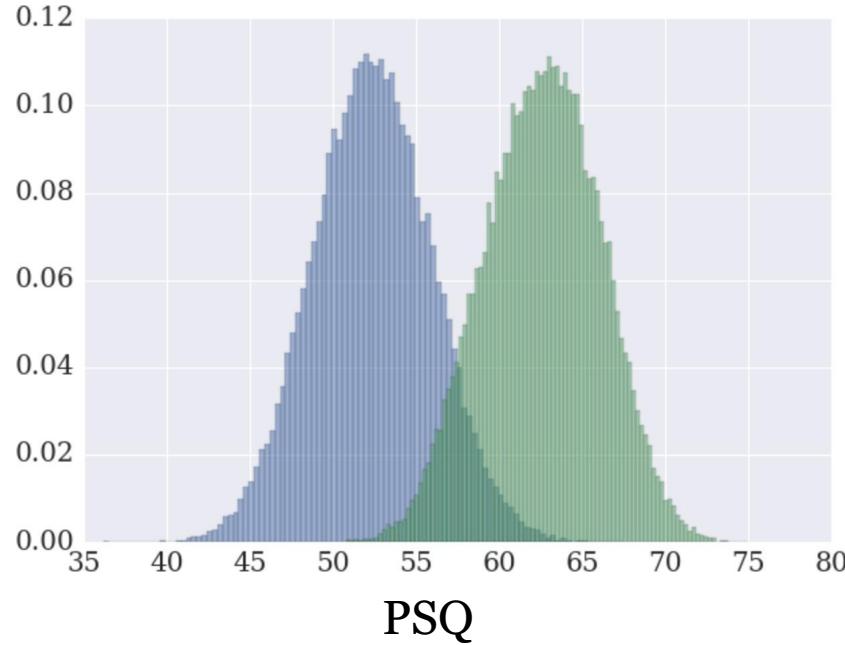
Non-overlapping Instances



Ground Truth

# PSQ – Humans vs Computers

	PSQ avg.	Seg Quality avg.	Det Quality avg.
<b>Humans</b>	<b>62.6%</b>	<b>83.9%</b>	<b>73.43%</b>
<b>Mask R-CNN + PSPNet</b>	<b>51.7%</b>	<b>81.0%</b>	<b>62.01%</b>

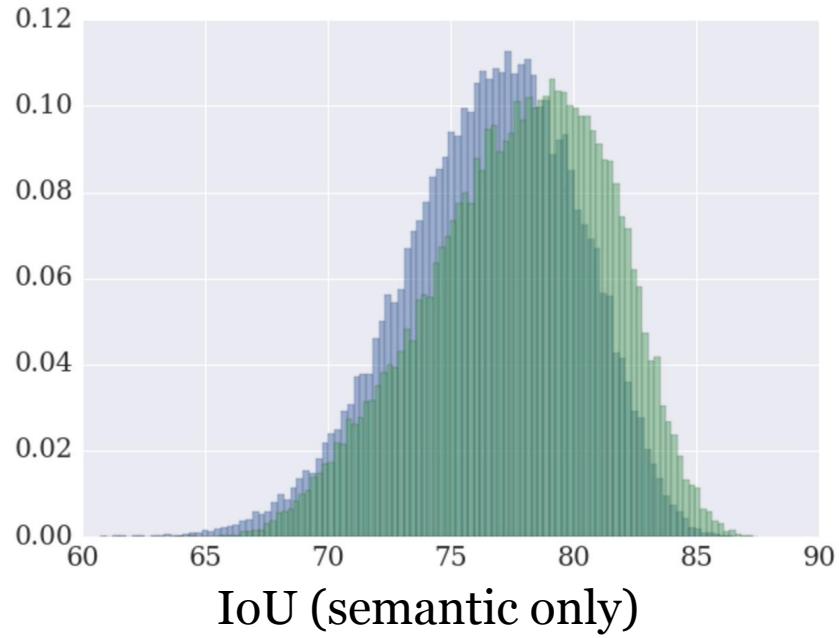


Humans

Heuristic combination of Mask R-CNN and PSPNet

# PSQ – Humans vs Computers

	PSQ avg.	Seg Quality avg.	Det Quality avg.
<b>Humans</b>	<b>62.6%</b>	<b>83.9%</b>	<b>73.43%</b>
<b>Mask R-CNN + PSPNet</b>	<b>51.7%</b>	<b>81.0%</b>	<b>62.01%</b>



Humans

Heuristic combination of Mask R-CNN and PSPNet

# PSQ – Humans vs Computers

<b>Cityscapes</b>	PQ	SQ	DQ	$PQ^{St}$	$PQ^{Th}$
human	$69.6^{+2.5}_{-2.7}$	$84.1^{+0.8}_{-0.8}$	$82.0^{+2.7}_{-2.9}$	$71.2^{+2.3}_{-2.5}$	$67.4^{+4.6}_{-4.9}$
machine	61.2	81.0	74.4	66.4	54.1
<b>ADE20k</b>	PQ	SQ	DQ	$PQ^{St}$	$PQ^{Th}$
human	$67.6^{+2.0}_{-2.0}$	$85.7^{+0.6}_{-0.6}$	$78.6^{+2.1}_{-2.1}$	$71.0^{+3.7}_{-3.2}$	$66.4^{+2.3}_{-2.4}$
machine	35.6	74.4	43.2	24.5	41.1
<b>Vistas</b>	PQ	SQ	DQ	$PQ^{St}$	$PQ^{Th}$
human	$57.7^{+1.9}_{-2.0}$	$79.7^{+0.8}_{-0.7}$	$71.6^{+2.2}_{-2.3}$	$62.7^{+2.8}_{-2.8}$	$53.6^{+2.7}_{-2.8}$
machine	38.3	73.6	47.7	41.8	35.7

# Outline

- Motivation
- Problem Definition
- Quality Evaluation
- Human Performance
- Humans vs Computers
- **Perspectives**

# Why solve it?

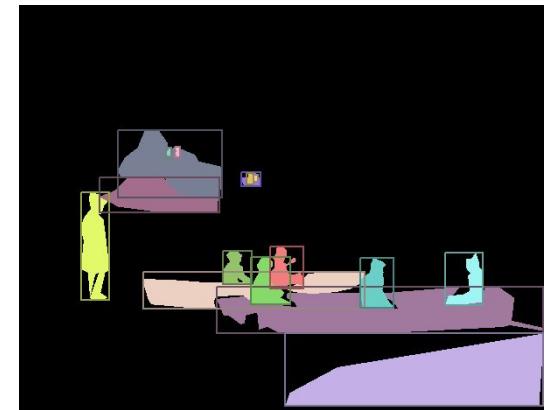


**Semantic Segmentation**

- per-pixel annotation
- simple accuracy measure
- instances  
indistinguishable



**Panoptic Segmentation**



**Object Detection/Seg**

- each object detected and segmented separately
- “stuff” is not segmented

# Why solve it?



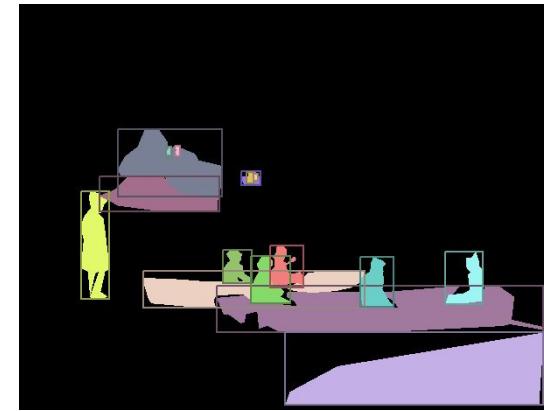
**Semantic Segmentation**

- per-pixel annotation
- simple accuracy measure
- instances  
**indistinguishable**

FCN 8s, Dilation8, DeepLab,  
PSPNet, RefineNet, U-Net,  
etc.



**Panoptic Segmentation**



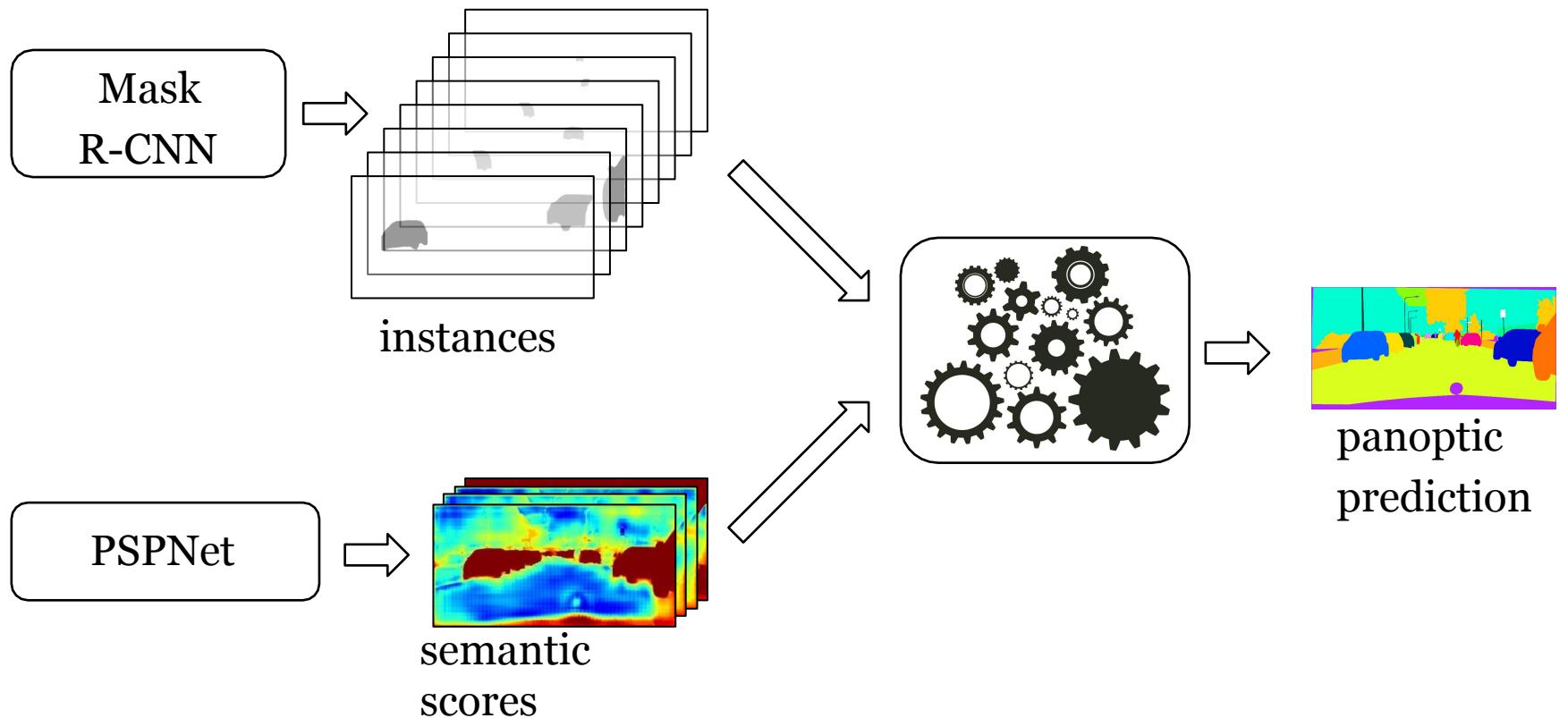
**Object Detection/Seg**

- each object detected and segmented separately
- “stuff” is not segmented

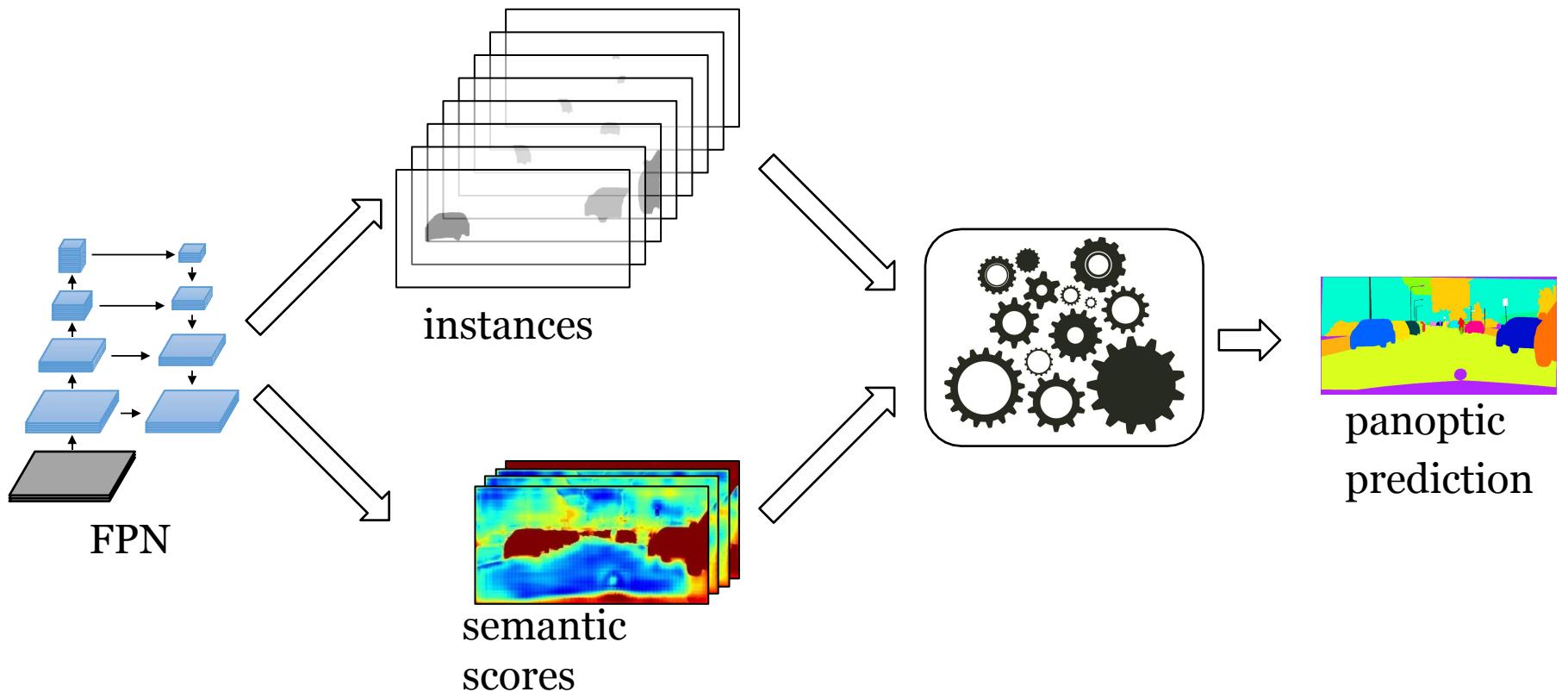
?

Fast/er R-CNN, DeepMask,  
SharpMask, Mask R-CNN,  
FCIS, YOLO, RetinaNet,  
FPN, etc.

# Why solve it?



# Why solve it?



# Why solve it?

