2018 Panoptic Challenge

Joint COCO and Mapillary Recognition Challenge Workshop

Sunday, September 9th, ECCV 2018

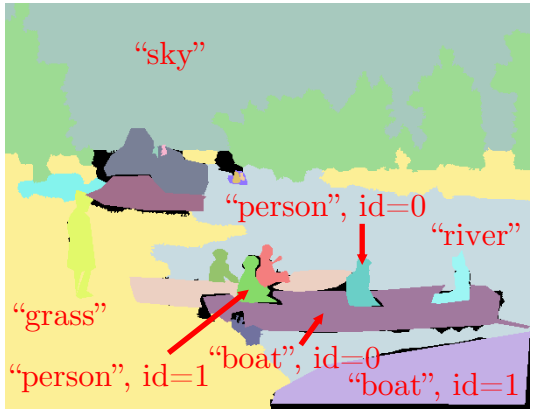
Alexander Kirillov, Facebook AI Research

Common Objects in Context

COCO Panoptic Dataset

2018 Panoptic Segmentation Dataset

COCO

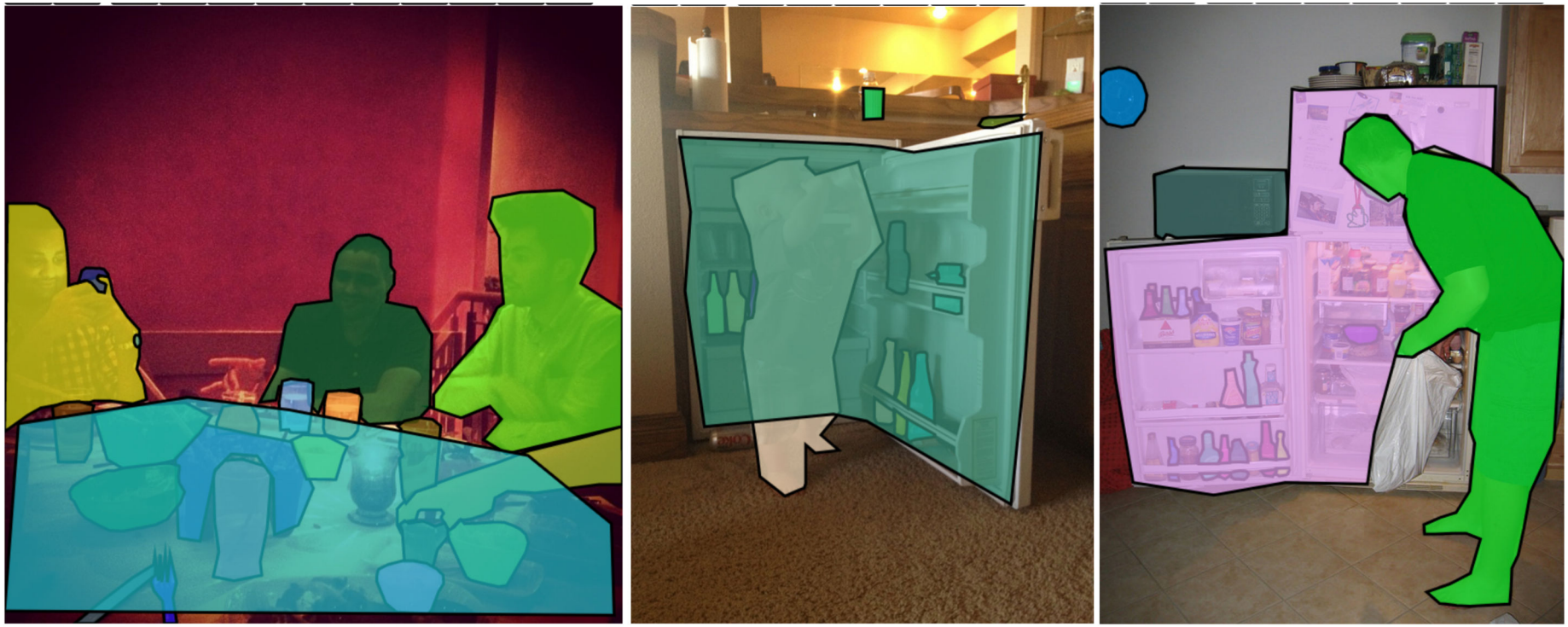


Panoptic COCO

COCO-stuff

- For each pixel i predict semantic label l and instance id z
- **no overlaps** between segments by design

2018 Panoptic Segmentation Dataset

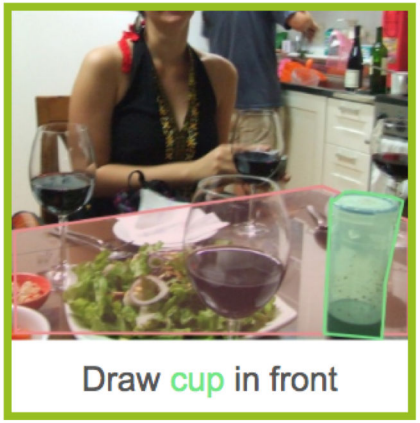


- COCO annotations have overlaps
- Most overlaps can be resolved automatically
- ~25k overlaps require manual resolution

2018 Panoptic Segmentation Dataset

Instructions:

In each row click on the image with better objects layout.



Draw cup in front

Draw table in front

Click to see more examples.

Task:

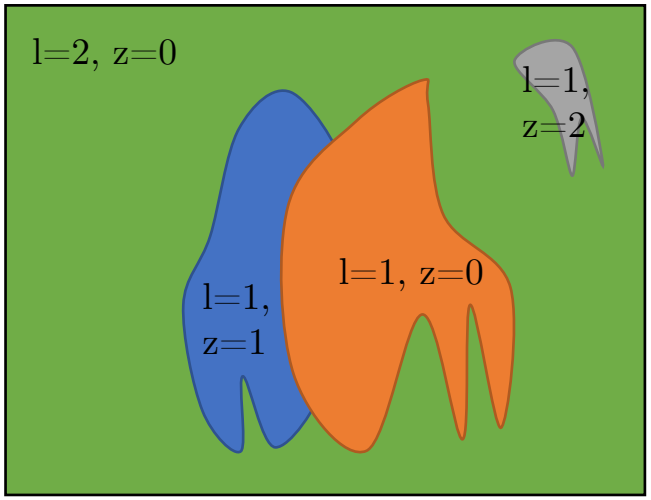
Draw person in front

Draw tennis racket in front

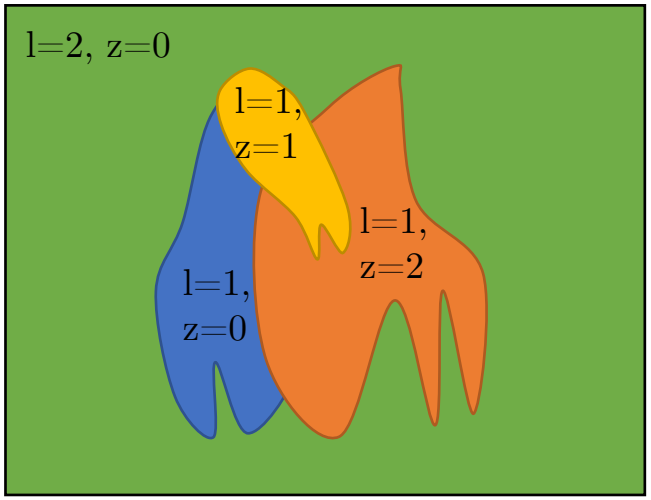
2018 Panoptic Segmentation Dataset

- train: 118k, val: 5k, test-dev: 20k, test-challenge: 20k
- 80 things categories, 53 stuff categories

Panoptic Quality Measure



Ground Truth

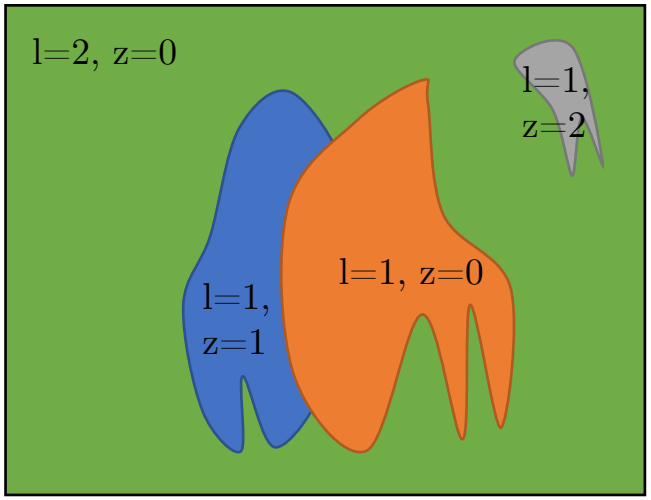


Prediction

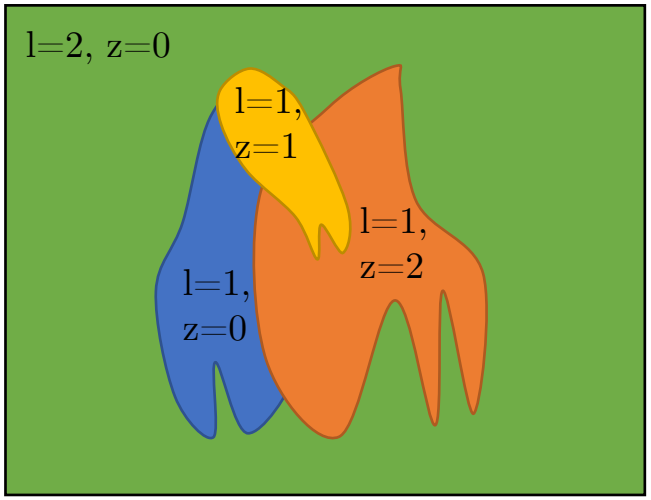
PQ Computation:

- Step 1: Matching
- Step 2: Calculation

Panoptic Quality (PQ): Matching



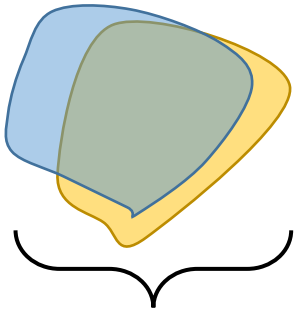
Ground Truth



Prediction

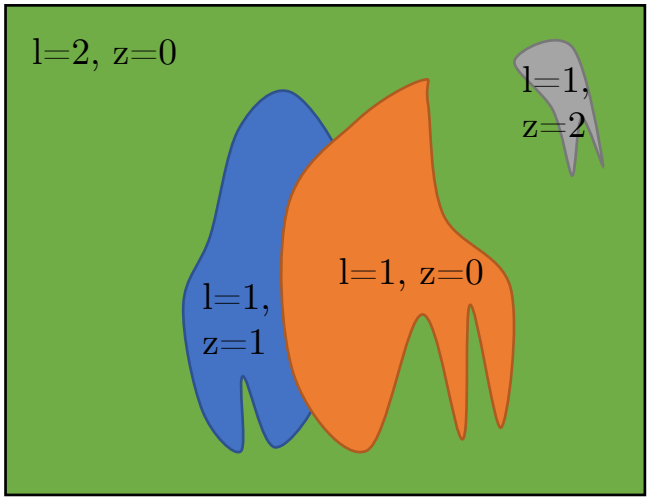
Theorem: For panoptic segmentation problem each ground truth segment can have at most one corresponding predicted segment with IoU greater than 0.5

Proof sketch:

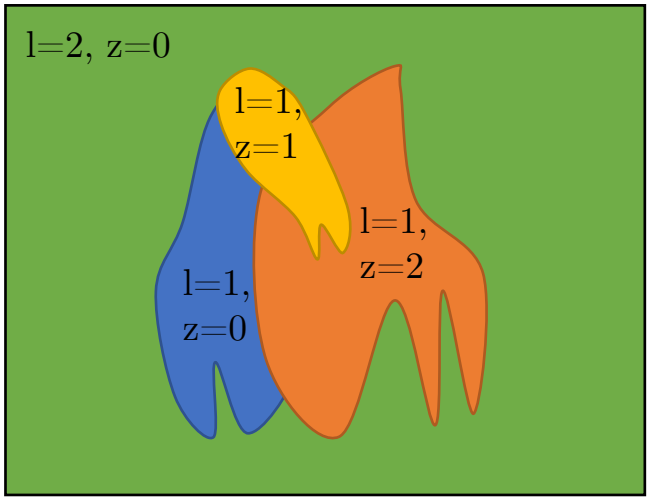
if 

then there is no other non overlapping object that has IoU > 0.5.

Panoptic Quality (PQ): Matching



Ground Truth



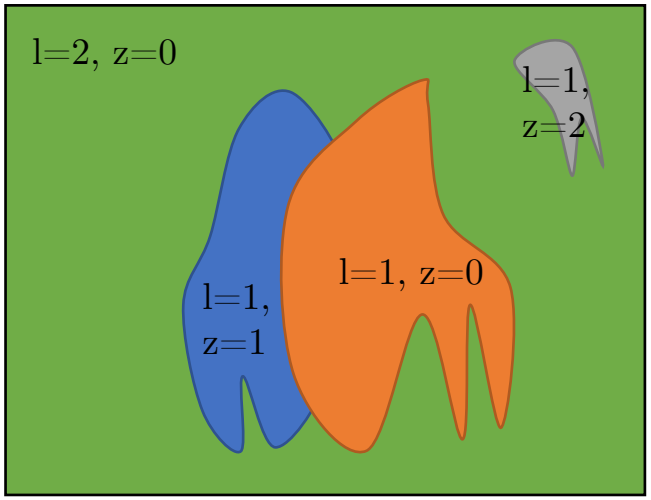
Prediction

$$\text{TP} = \{(\boxed{}, \boxed{}), (\boxed{}, \boxed{})\}$$

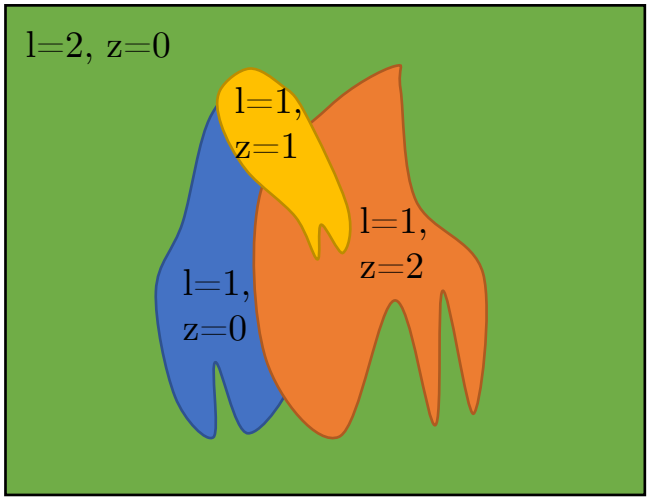
$$\text{FP} = \{\boxed{}\}$$

$$\text{FN} = \{\boxed{}\}$$

Panoptic Quality (PQ):Calculation



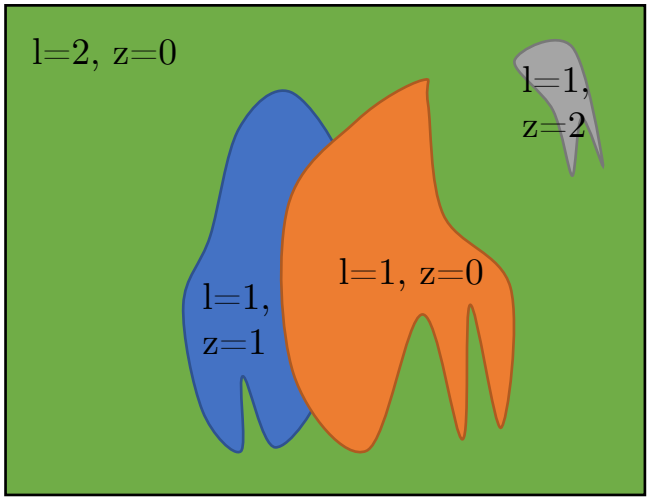
Ground Truth



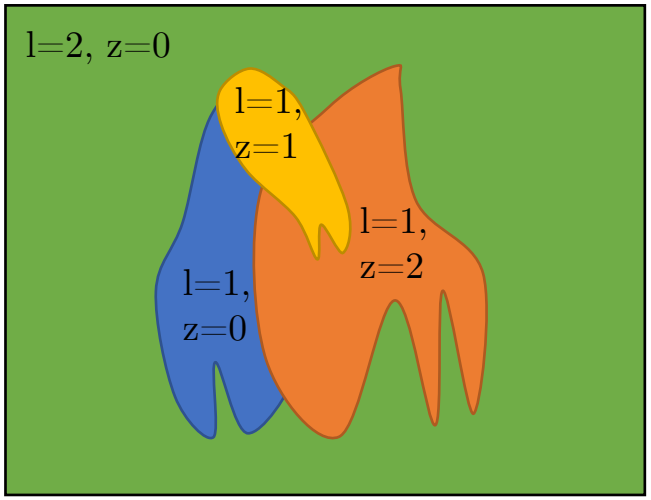
Prediction

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

Panoptic Quality (PQ):Calculation



Ground Truth



Prediction

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

COCO Panoptic Metrics

```
Average Panoptic Metrics:
   PQ                  % Panoptic Quality (primary challenge metric)
   SQ                  % Segmentation Quality component of PQ
   RQ                  % Recognition Quality component of PQ
Panoptic Metrics for Things Categories:
   PQ^Th               % PQ for things categories only
   SQ^Th               % SQ for things categories only
   RQ^Th               % RQ for over things categories only
Panoptic Metrics for Stuff Categories:
   PQ^St               % PQ for stuff categories only
   SQ^St               % SQ for stuff categories only
   RQ^St               % RQ for stuff categories only
```

COCO Annotations Consistency

5000 COCO images were annotated independently twice

COCO Annotations Consistency

5000 COCO images were annotated independently twice

	PQ	SQ	RQ
All	53.5	82.6	63.9
Things	57.8	81.4	69.7
Stuff	47.1	84.3	55.2

- Crowd sourced annotations are very noisy

COCO Annotations Consistency

5000 COCO images were annotated independently twice

	PQ	SQ	RQ
All	53.5	82.6	63.9
Things	57.8	81.4	69.7
Stuff	47.1	84.3	55.2

	PQ	SQ	RQ
Small	25.2	62.1	32.8
Medium	53.5	81.7	64.6
Large	69.6	87.5	78.3

- Crowd sourced annotations are very noisy
- Annotations are highly inconsistent for small objects

COCO Annotations Consistency

Annotations Consistency $<$ Human Performance

COCO Annotations Consistency

Annotations Consistency $<$ Human Performance

$$
\begin{aligned}
\text{real GT} &= [0, 0, 0, 0, 0, 1, 1, 1, 1, 1] \\
\text{noisy annotator 1} &= [0, 0, 0, 0, 1, 1, 1, 1, 1, 1] \\
\text{noisy annotator 2} &= [0, 0, 0, 0, 0, 0, 1, 1, 1, 1]
\end{aligned}
$$

COCO Annotations Consistency

Annotations Consistency $<$ Human Performance

$$
\begin{aligned}
\text{real GT} &= [0, 0, 0, 0, 0, 1, 1, 1, 1, 1] \\
\text{noisy annotator 1} &= [0, 0, 0, 0, 1, 1, 1, 1, 1, 1] \\
\text{noisy annotator 2} &= [0, 0, 0, 0, 0, 0, 1, 1, 1, 1]
\end{aligned}
$$

$$
\begin{aligned}
\text{Accuracy(real GT, annotator\_1)} &= 0.9 \\
\text{Accuracy(real GT, annotator\_2)} &= 0.9 \\
\text{Accuracy(annotator\_1, annotator\_2)} &= 0.8
\end{aligned}
$$

COCO Annotations Consistency

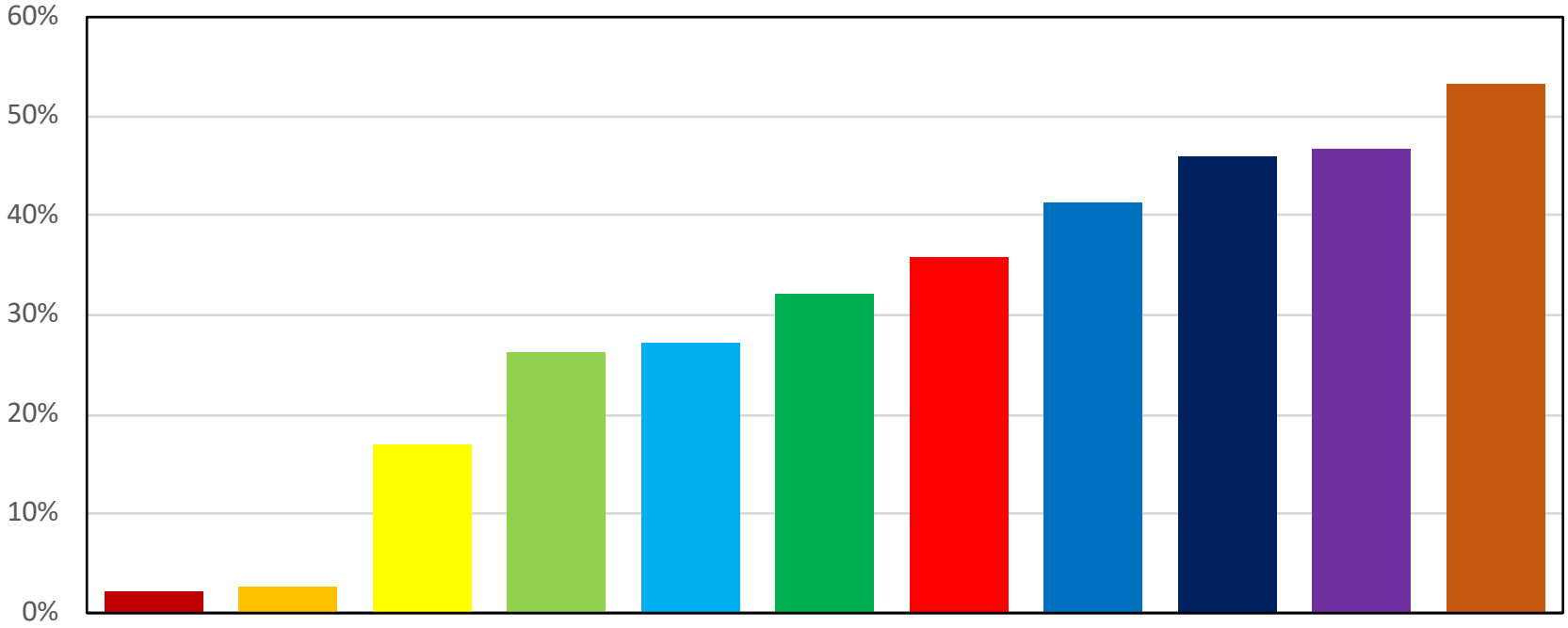
Annotations Consistency $<$ Human Performance

$$
\begin{aligned}
\text{real GT} &= [0, 0, 0, 0, 0, 1, 1, 1, 1, 1] \\
\text{noisy annotator 1} &= [0, 0, 0, 0, 1, 1, 1, 1, 1, 1] \\
\text{noisy annotator 2} &= [0, 0, 0, 0, 0, 0, 1, 1, 1, 1]
\end{aligned}
$$

$$
\begin{aligned}
\text{Accuracy(real GT, annotator\_1)} &= 0.9 \\
\text{Accuracy(real GT, annotator\_2)} &= 0.9 \\
\text{Accuracy(annotator\_1, annotator\_2)} &= 0.8
\end{aligned}
$$

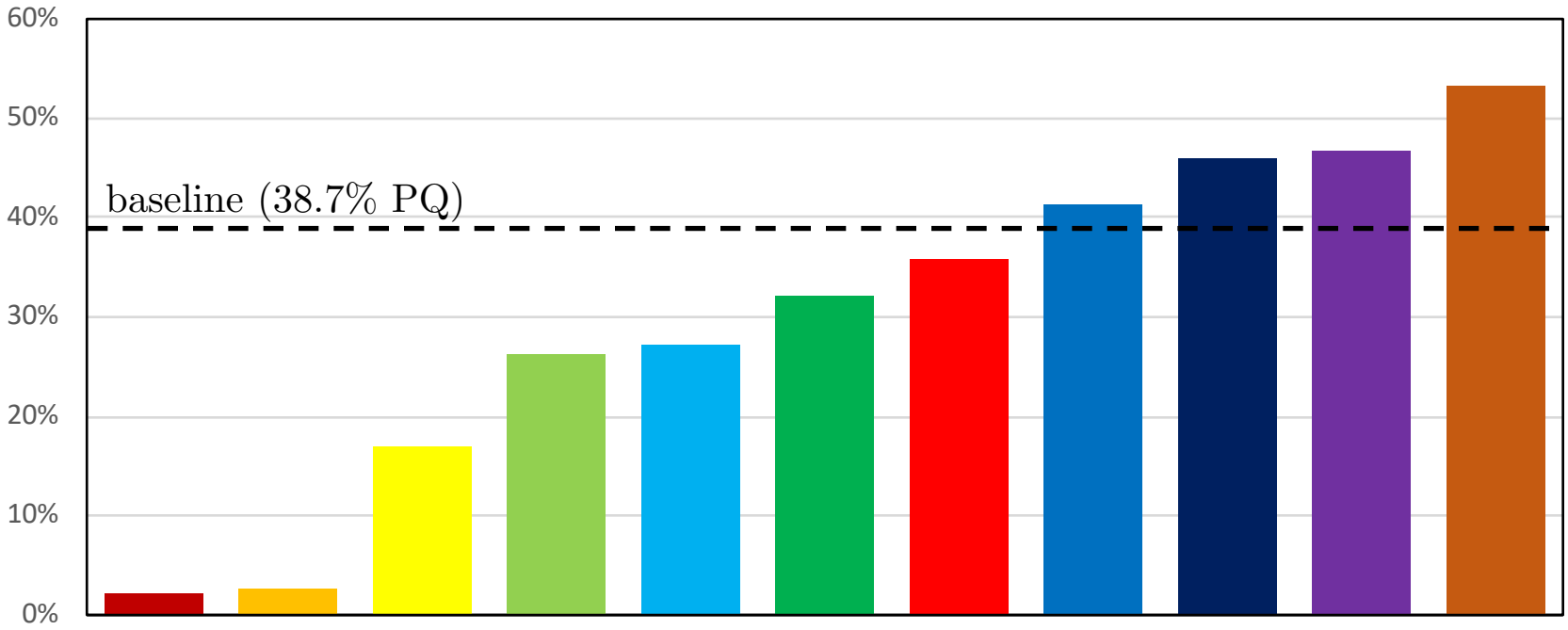
$$
\text{Accuracy(real GT, ideal annotator)} = ?
$$

2018 COCO Panoptic Challenge



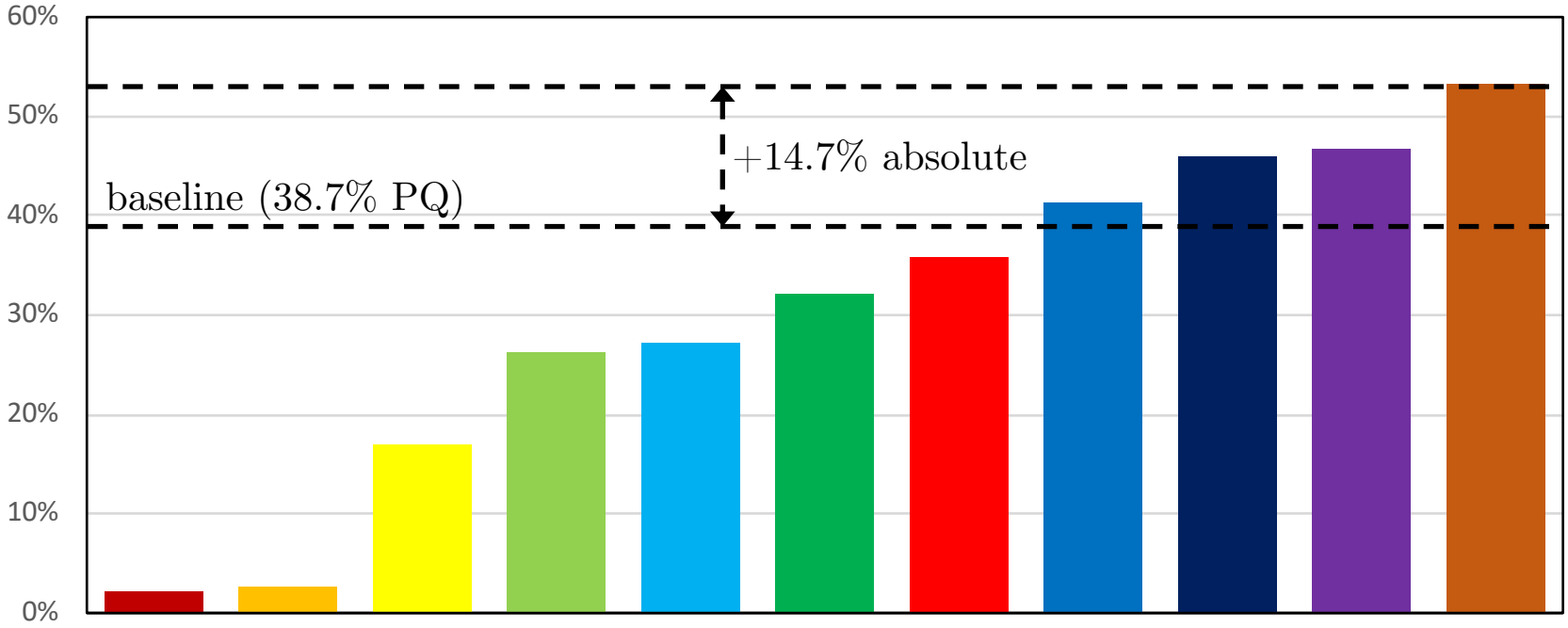
- 11 teams joined the competition

2018 COCO Panoptic Challenge



- 11 teams joined the competition
- 4 teams achieved better performance than the baseline (RN50 Mask R-CNN + RN50 FPN-FCN)

2018 COCO Panoptic Challenge



- 11 teams joined the competition
- 4 teams achieved better performance than the baseline (RN50 Mask R-CNN + RN50 FPN-FCN)

Summary of Findings

2018 COCO Panoptic Challenge Take-aways

- All submission above the baseline combined the outputs of two separate networks for stuff and things

Summary of Findings

2018 COCO Panoptic Challenge Take-aways

- All submission above the baseline combined the outputs of two separate networks for stuff and things
- Best submission showed better PQ for things categories than the human consistency experiment

	PQ	SQ	RQ	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{St}	SQ^{St}	RQ^{St}
Human Consistency	53.5	82.6	**63.9**	57.8	81.4	69.7	**47.1**	**84.3**	**55.2**
Megvii (Face++)	**53.8**	**83.4**	63.6	**62.8**	**85.7**	**73.1**	40.2	80	49.2

Summary of Findings

2018 COCO Panoptic Challenge Take-aways

- All submission above the baseline combined the outputs of two separate networks for stuff and things
- Best submission showed better PQ for things categories than the human consistency experiment

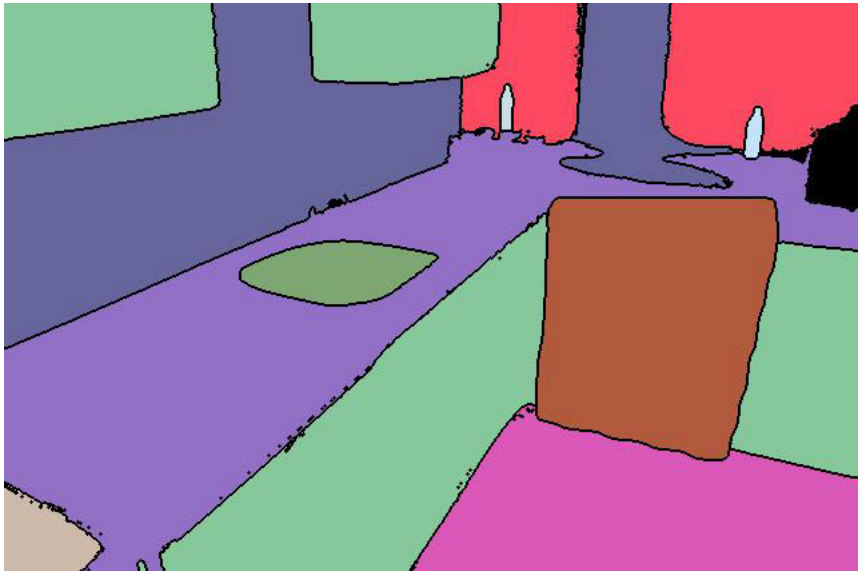
Things	PQ^{Th}	SQ^{Th}	RQ^{Th}	all |TP|	all |FP|	all |FN|	Precision	Recall
Human Consistency	57.8	81.4	69.7	24890	8860	9628	72.6%	**69.5%**
Megvii (Face++)	**62.8**	**85.7**	**73.1**	24205	4929	10313	**81.1%**	67.1%

- The result suggests ability to learn models with low noise level from large-scale noisy data
- Accuracy of the test set ground truth needs to be improved in the future

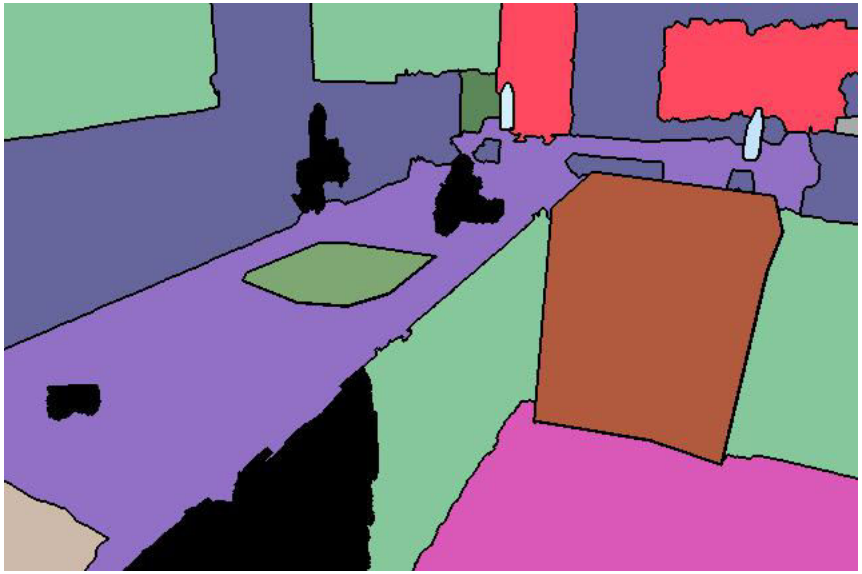
Prediction Examples

Image

Prediction
Megvii (Face++)



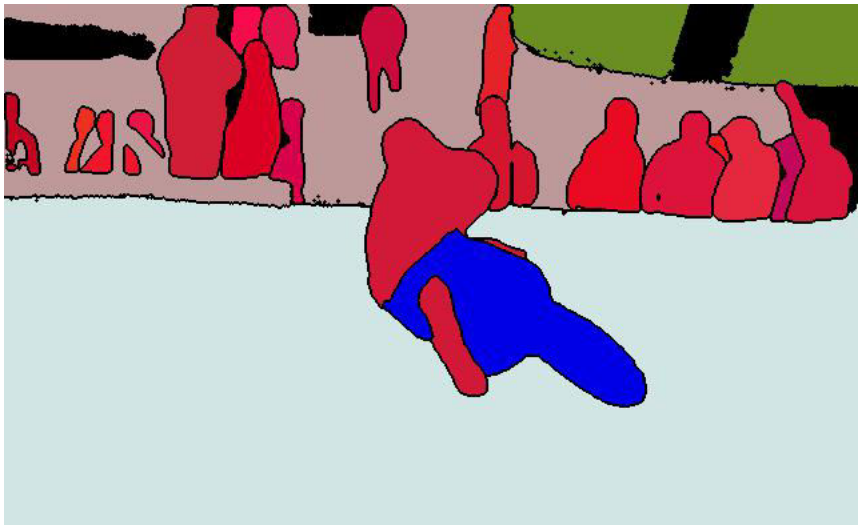
GT



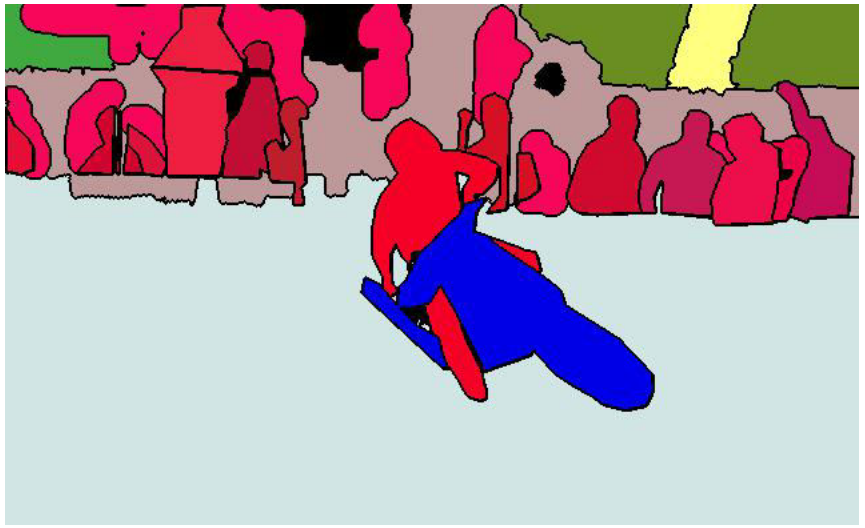
Prediction Examples

Image

Prediction
Megvii (Face++)



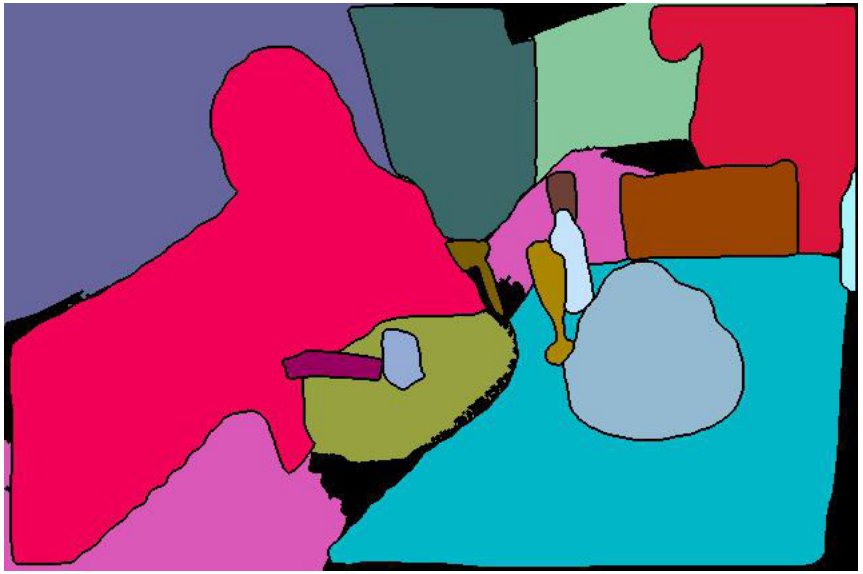
GT



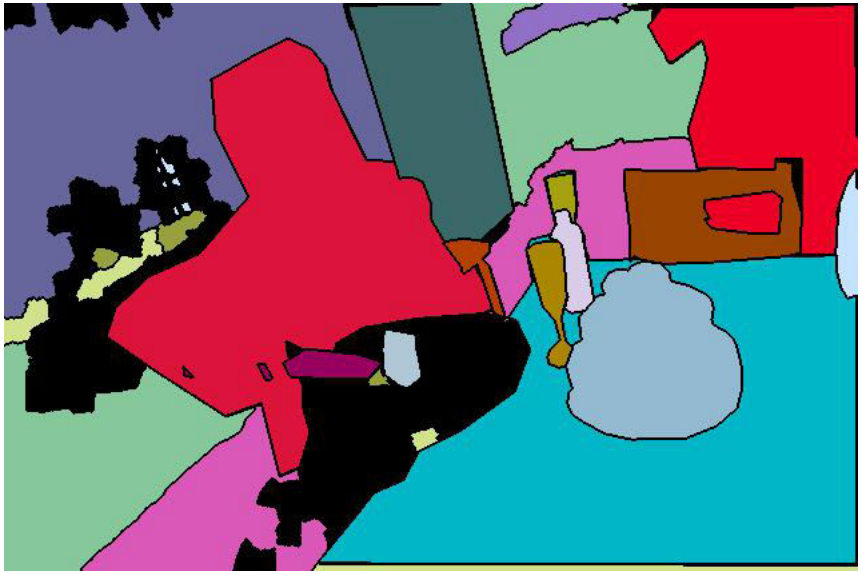
Prediction Examples

Image

Prediction
Megvii (Face++)



GT

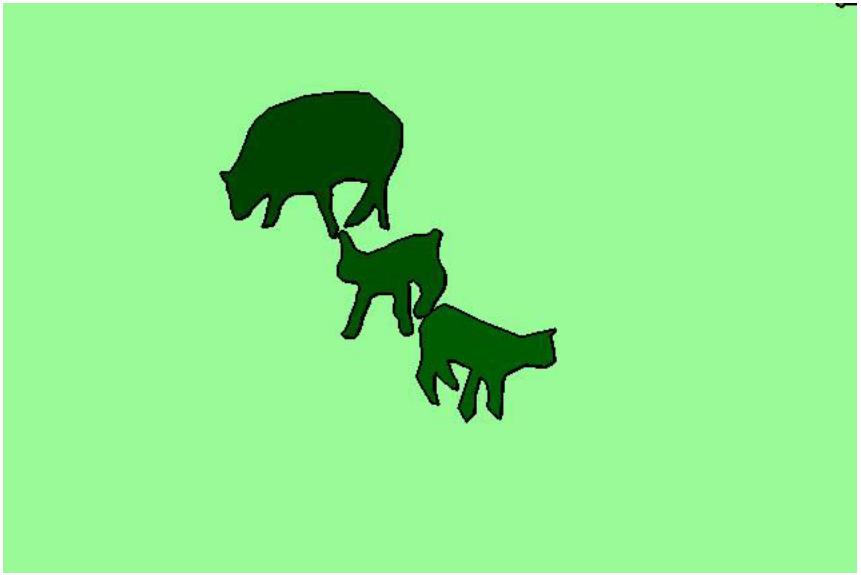


Prediction Examples

Image

Prediction
Megvii (Face++)

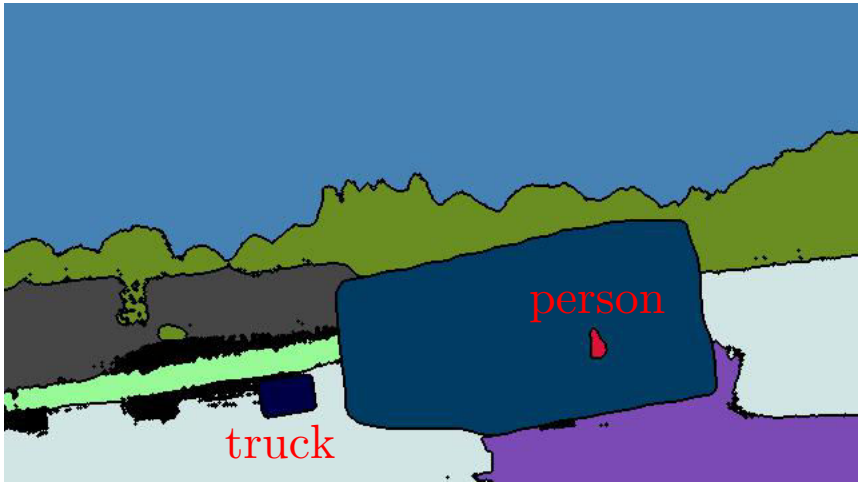
GT



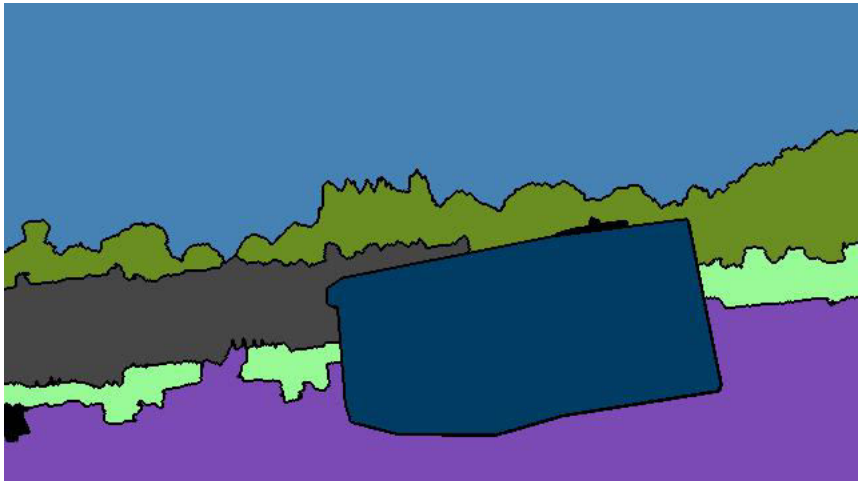
Prediction Examples

Image

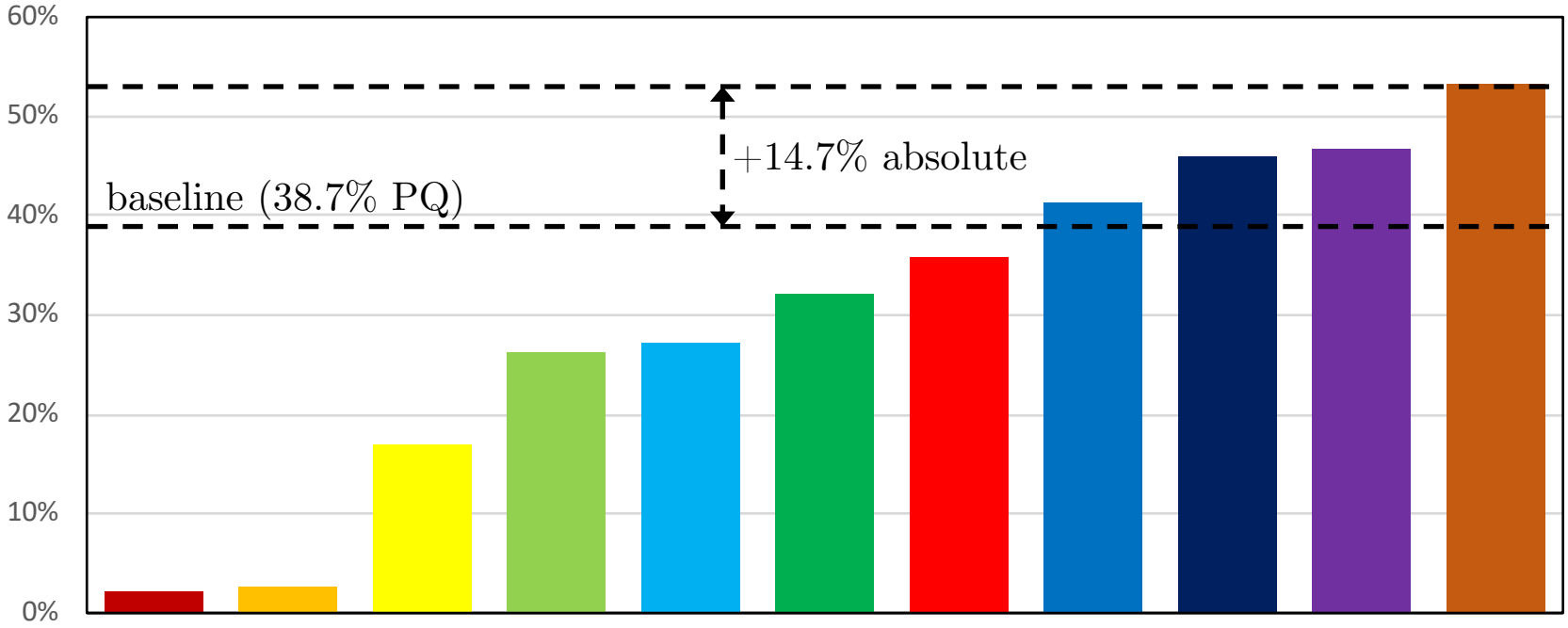
Prediction
Megvii (Face++)



GT

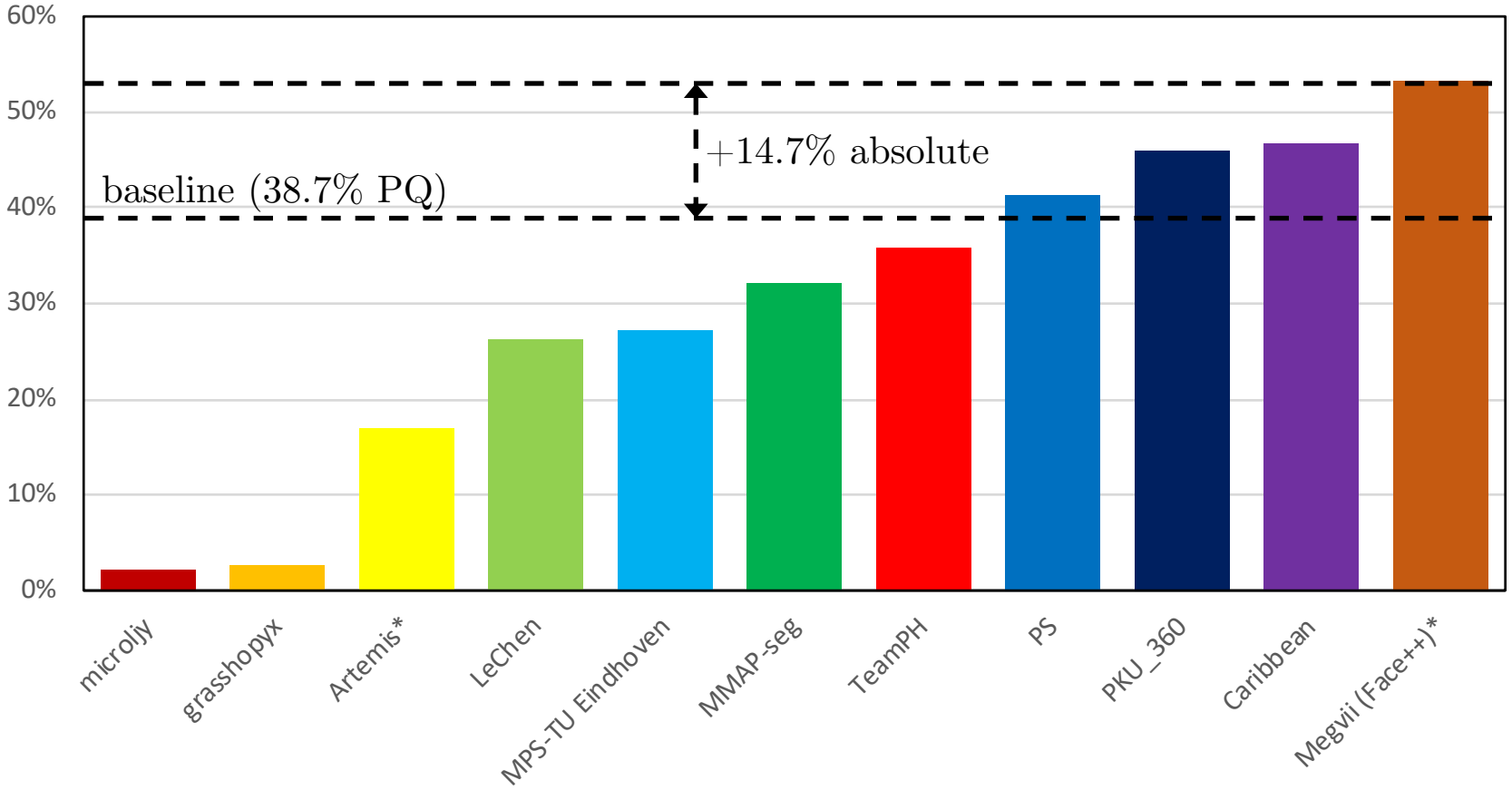


2018 COCO Panoptic Challenge



- 11 teams joined the competition
- 4 teams achieved better performance than the baseline (RN50 Mask R-CNN + RN50 FPN-FCN)

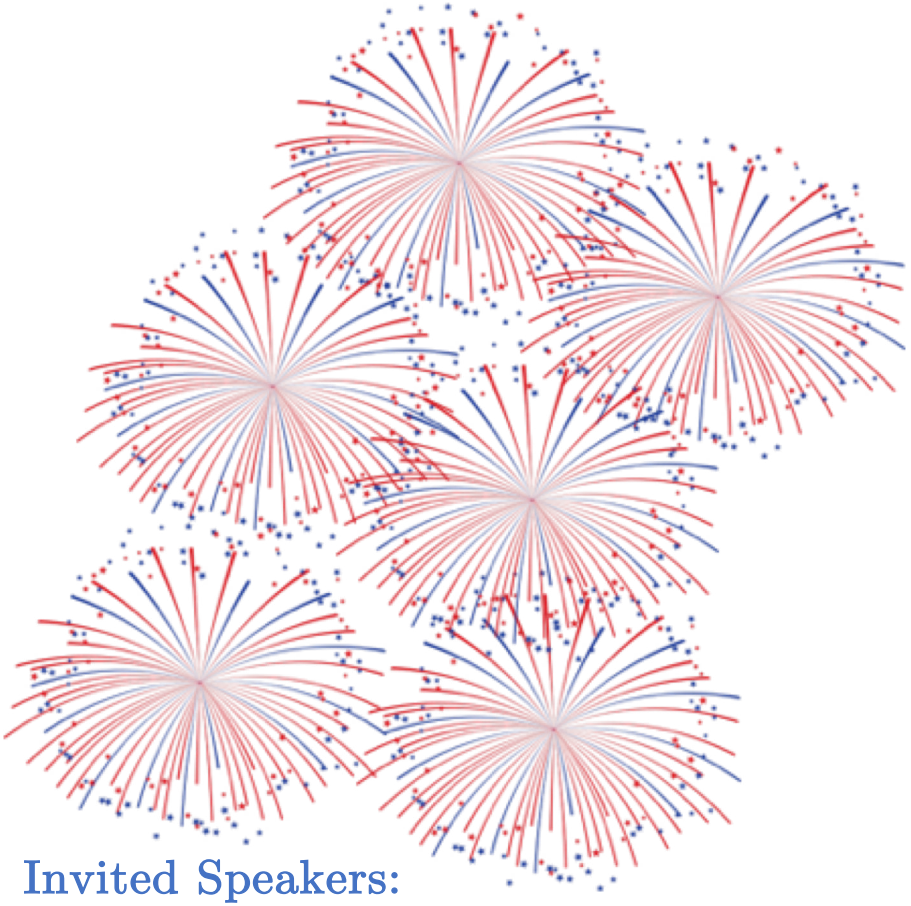
2018 COCO Panoptic Challenge



- 11 teams joined the competition
- 4 teams achieved better performance than the baseline (RN50 Mask R-CNN + RN50 FPN-FCN)

*External segmentation datasets were used

2018 COCO Panoptic Challenge



Invited Speakers:

Team Megvii / (9:15am – 9:45am)

Team Caribbean / (11:30am – 11:45am)

Team PKU_360 / (11:45am – 12:00pm)

Team	Position
Megvii (Face++)	1st
Caribbean	2nd
PKU_360	3rd