北京大學
PEKING UNIVERSITY

Scene Overlap Graph for Panoptic Segmentation

COCO Challenge 2019 Panoptic Segmentation Track

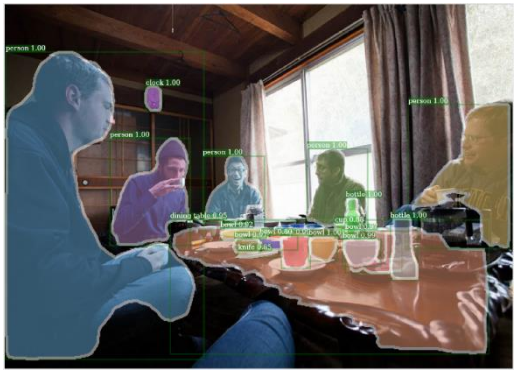
Team: PKU_ZERO

Yibo Yang

with Xia Li, HongYang Li, Tiancheng Shen, Yudong Liu & Zhouchen Lin
Peking University

Introduction

Image



Instance Segmentation

↓

Permit overlaps

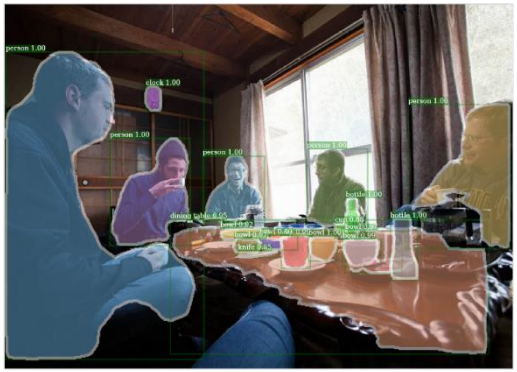
Panoptic Segmentation

↓

no overlapping segments

Introduction

Image



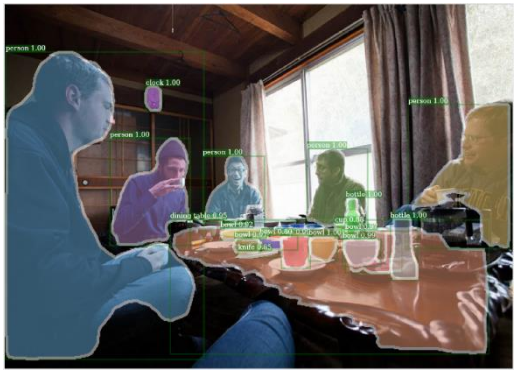
Instance Segmentation

Panoptic Segmentation

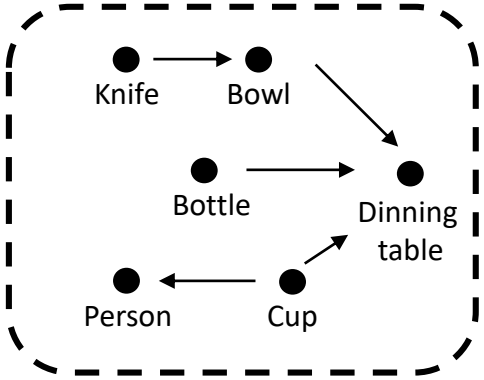
How to solve the overlap problem?

- Heuristic rules
- Panoptic head to predict
- Our method: Explicitly modeling overlap relations

Introduction



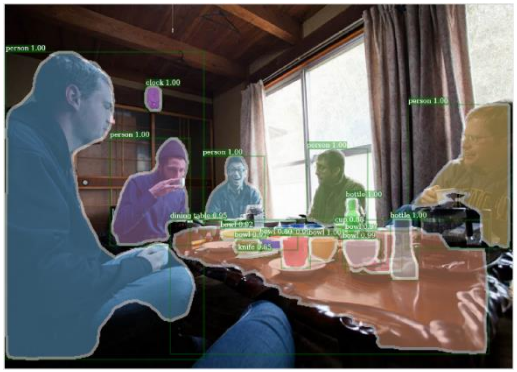
Instance Segmentation



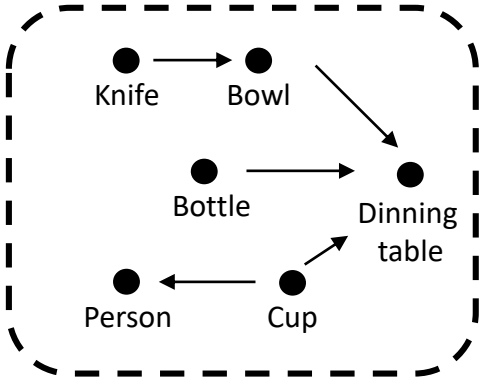
Overlap relations

Panoptic Segmentation

Introduction



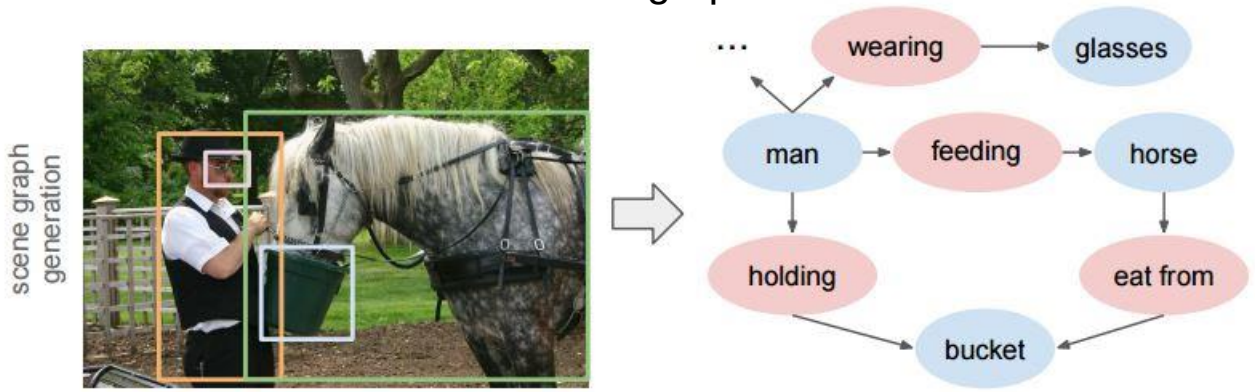
Instance Segmentation



Overlap relations

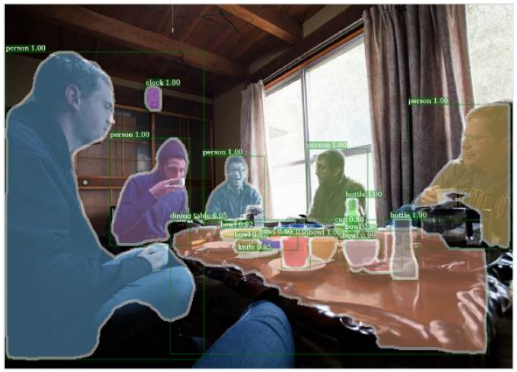
Panoptic Segmentation

A scene graph

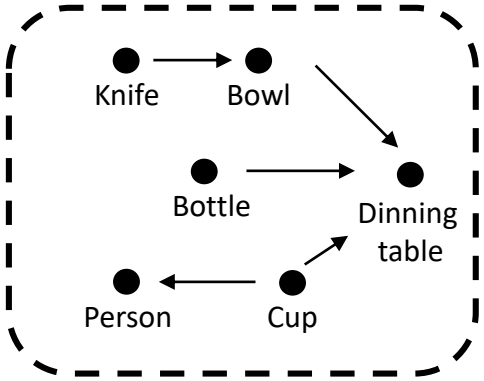


Relations: wear, feed, hold, behind, under …

Introduction



Instance Segmentation



Overlap relations

Panoptic Segmentation

A *simplified* scene graph

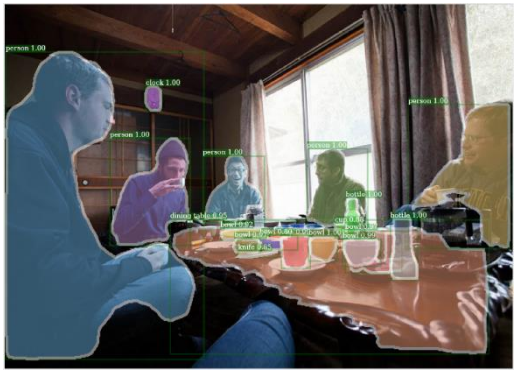
Instance i cover j

No relation

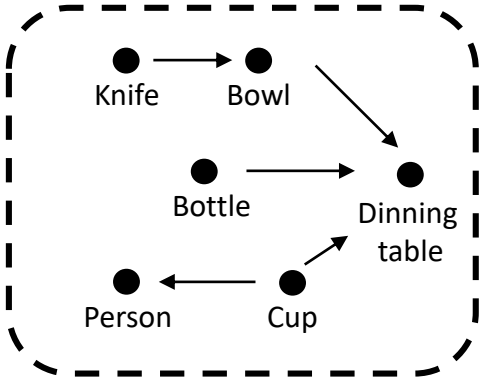
Instance i is covered by j

Scene Overlap Graph

Introduction



Instance Segmentation



Overlap relations

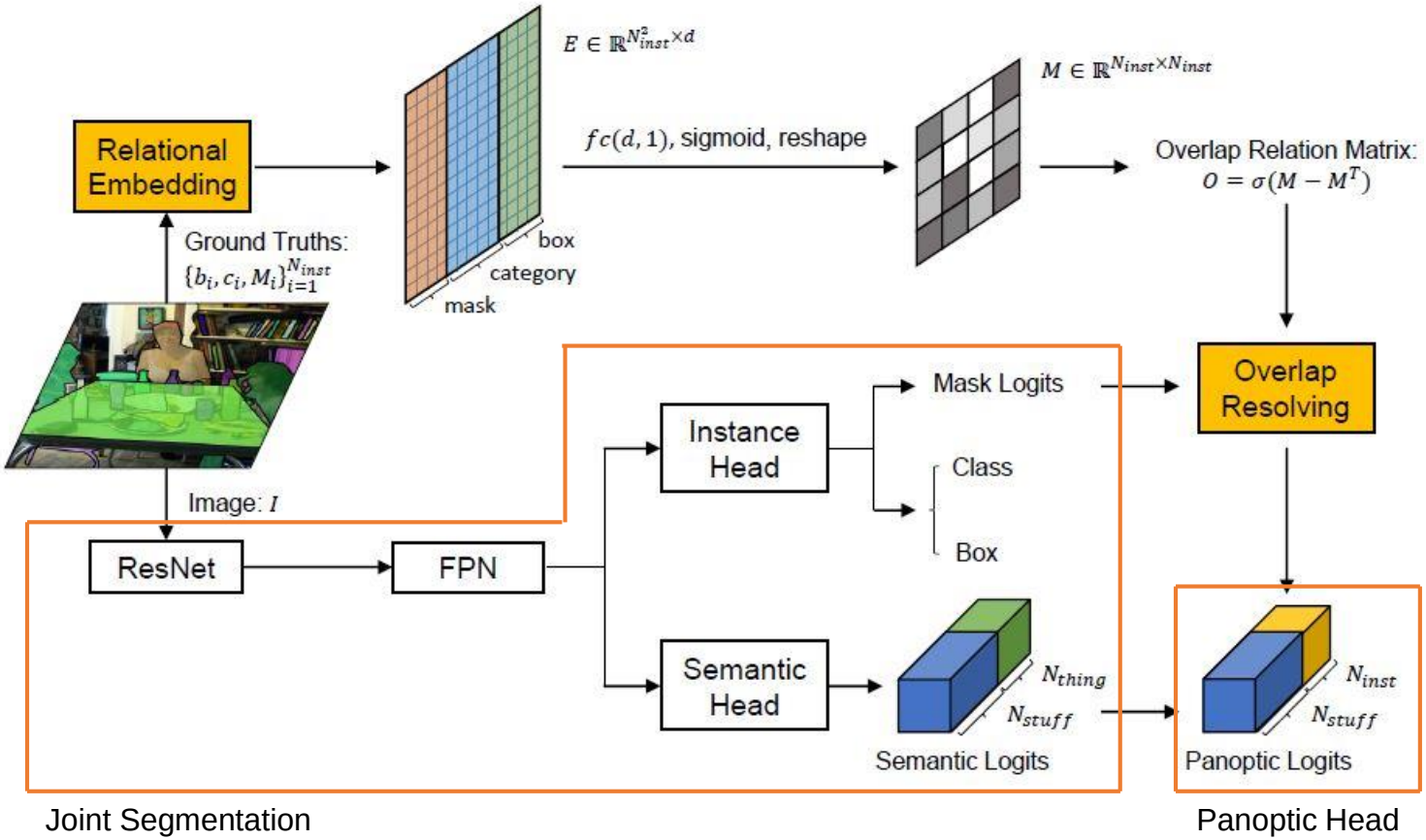
Panoptic Segmentation

A *simplified* scene graph

Problem:

Different from scene graph parsing tasks, panoptic segmentation does not offer annotations of object relations, or depth information, so overlap relations cannot be trained with direct supervision.

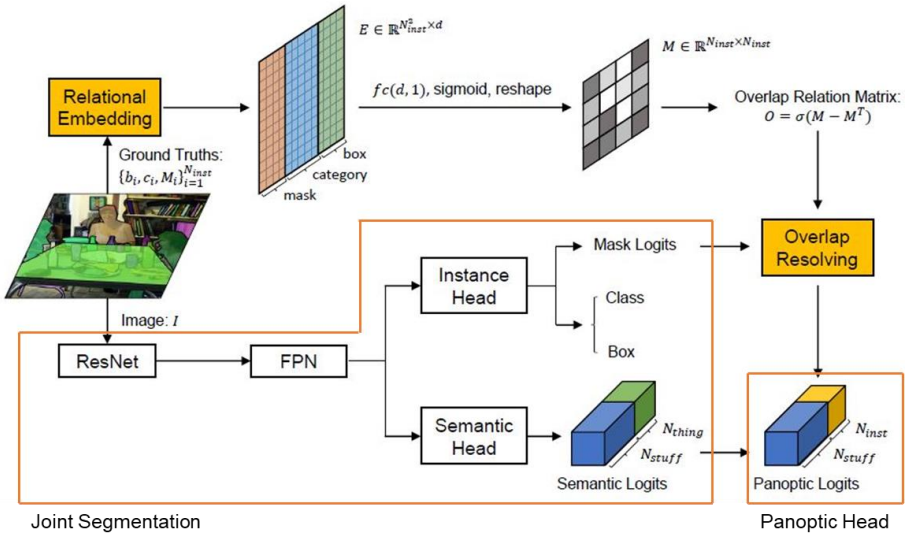
SOGNet


Relational Embedding
$E \in \mathbb{R}^{N_{inst}^2 \times d}$
box
category
mask
$fc(d, 1)$, sigmoid, reshape
$M \in \mathbb{R}^{N_{inst} \times N_{inst}}$
Overlap Relation Matrix:
$O = \sigma(M - M^T)$
Ground Truths:
$\{b_i, c_i, M_i\}_{i=1}^{N_{inst}}$
Image: I
ResNet
FPN
Instance Head
Mask Logits
Class
Box
Semantic Head
N_{thing}
N_{stuff}
Semantic Logits
Overlap Resolving
N_{inst}
N_{stuff}
Panoptic Logits
Joint Segmentation
Panoptic Head

SOGNet

Joint Segmentation

- Instance head: Mask R-CNN
- Semantic head:



FPN feature maps first go through three deformable 3x3 convolution layers, and then are up-sampled to the 1=4 scale. Finally, they are concatenated to generate the per-pixel category prediction.
Semantic branch is supervised with both stuff and thing classes, and predict all categories.

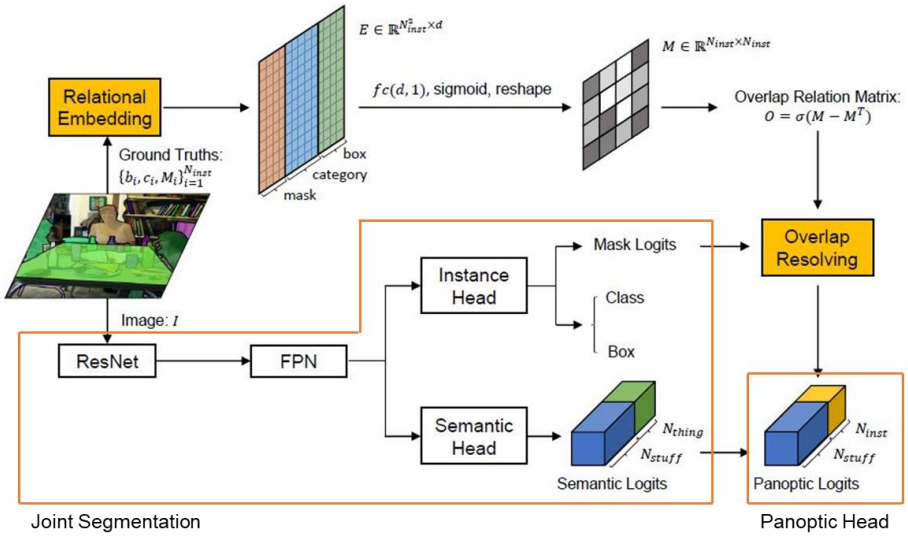
SOGNet

Relational Embedding

Ground truth $\{b_i, c_i, M_i\}_{i=1}^{N_{inst}}$

$b_i \in R^4, c_i \in R^{80}$

we resize the values inside box b_i from M_i as 28x28 to have $m_i \in R^{784}$.



$$E_{i|j}^{(c)} = P^T \left(\sigma(V^T c_i) \circ \sigma(U^T c_j)\right), \quad (1)$$

$$E^{(c)} = \left[E_{1|1}^{(c)}, E_{1|2}^{(c)}, \cdots, E_{N_{inst}|N_{inst}}^{(c)}\right]^T \in \mathcal{R}^{N_{inst}^2 \times d_c}, \quad (2)$$

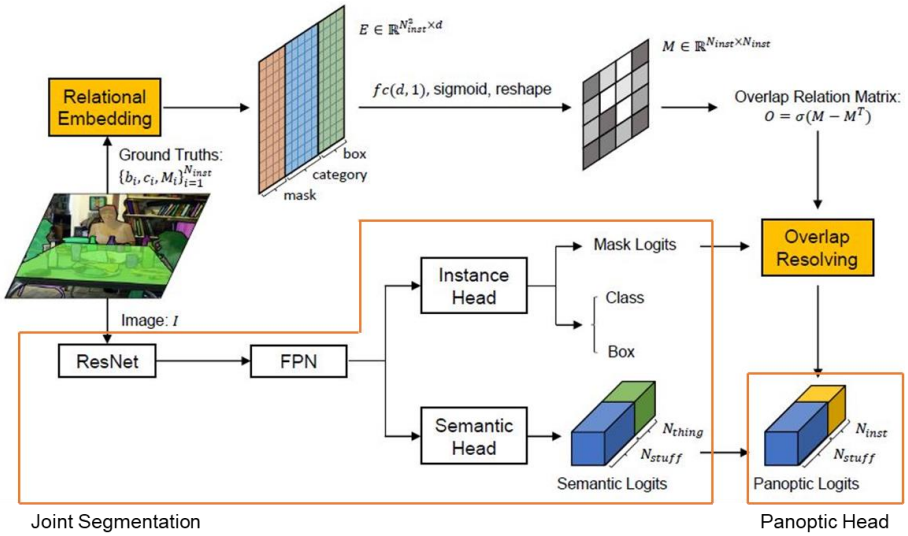
$$E_{i|j}^{(b)} = K^T \left(\frac{x_i - x_j}{w_j}, \frac{y_i - y_j}{h_j}, \log\left(\frac{w_i}{w_j}\right), \log\left(\frac{h_i}{h_j}\right)\right)^T, \quad (3)$$

$$E = [E^{(m)}, E^{(c)}, E^{(b)}] \in \mathcal{R}^{N_{inst}^2 \times d}, \quad (4)$$

SOGNet

Encode Overlap Relations

$$E \in R^{N_{inst}^2 \times d} \xrightarrow[\text{sigmoid}]{fc(d,1)} M \in R^{N_{inst} \times N_{inst}}$$



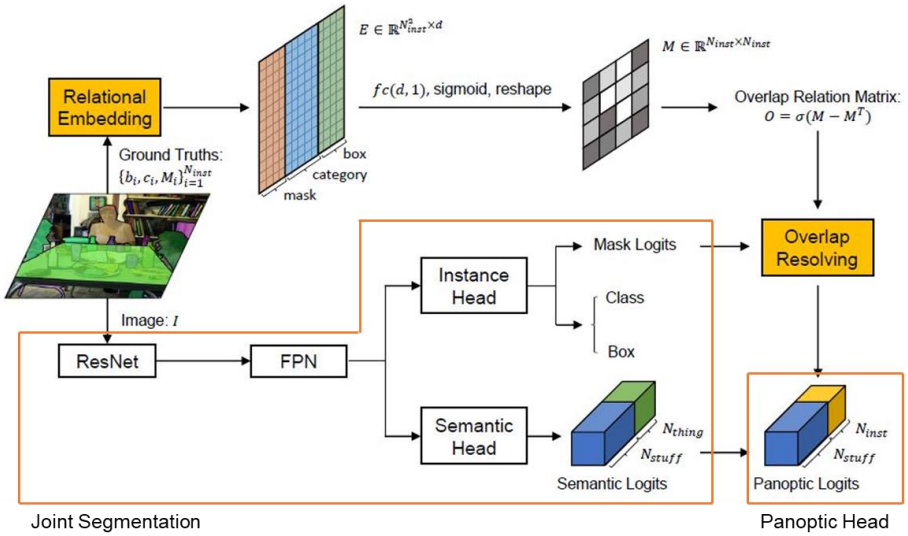
$$O = \sigma(M - M^T) \in \mathcal{R}^{N_{inst} \times N_{inst}}, \quad (5)$$

σ: ReLU actiavtion

$O_{ij} > 0$: instance i is covered by j, $O_{ji} > 0$

$O_{ij} = O_{ji} = 0$: no overlap between (i, j)

SOGNet



Overlap Resolving

Overlap between (i, j) :

$$A_i' = A_i - A_i \circ [s(A_i) \circ s(A_j)] O_{ij}, \quad (6)$$

Considering the overlap relations of all the other instances on i:

$$A_i' = A_i - A_i \circ s(A_i) \circ \sum_{j=1}^{N_{inst}} s(A_j) O_{ij}, \quad (7)$$

Computational step:

$$\mathcal{A}' = \mathcal{A} - \mathcal{A} \circ s(\mathcal{A}) \circ \left(s(\mathcal{A}) \times_3 O^T\right), \quad (8)$$

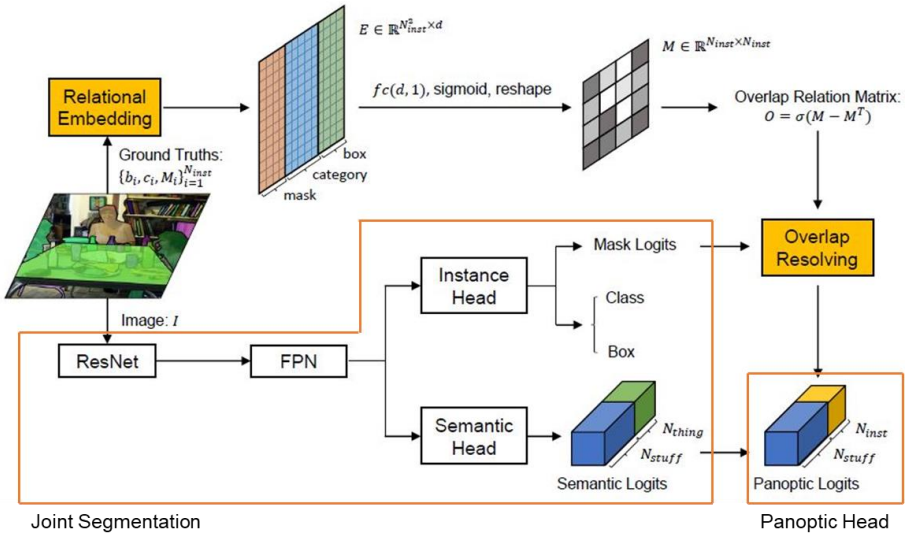
where $\mathcal{A} = [A_1, A_2, \cdots, A_{N_{inst}}] \in \mathcal{R}^{H \times W \times N_{inst}}$,

$\times_3$: reshape $s(\mathcal{A})$ as $R^{HW \times N_{inst}}$ for inner product with O^T, and then return to $R^{H \times W \times N_{inst}}$.

SOGNet

Panoptic Head

$$\text{Panoptic Head 1}: \quad Z_i = X_i + A'_i,$$
$$\text{Panoptic Head 2}: \quad Z_i = k \cdot X_i \circ s(A'_i) + A'_i,$$



k is a factor to balance the numerical difference between semantic output values and mask logtis.

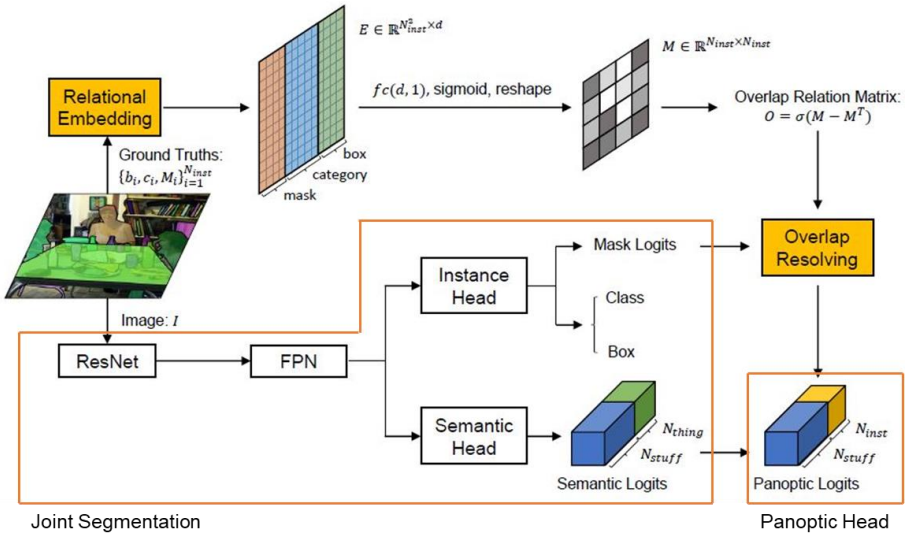
X_i: taking the values inside its ground box B_i from the channel corresponding to its ground truth category C_i, and padding zeros outside the box.

$\mathcal{L}_{panoptic}$: standard cross entropy loss for instance id classification.

SOGNet

Relation Loss

Use the ground truth binary masks $\{M_i\}_{i=1}^{N_{inst}}$ to infer whether to instances have overlaps or not:



$$R_{ij} = \mathbb{1}\left[\frac{|M_i \circ M_j|}{\min\{|M_i|, |M_j|\}} \geq 0.1\right], \quad i \neq j, \tag{11}$$

where $|\cdot|$ calculates the area of a binary mask. With the symmetric matrix R, we can introduce the relation loss:

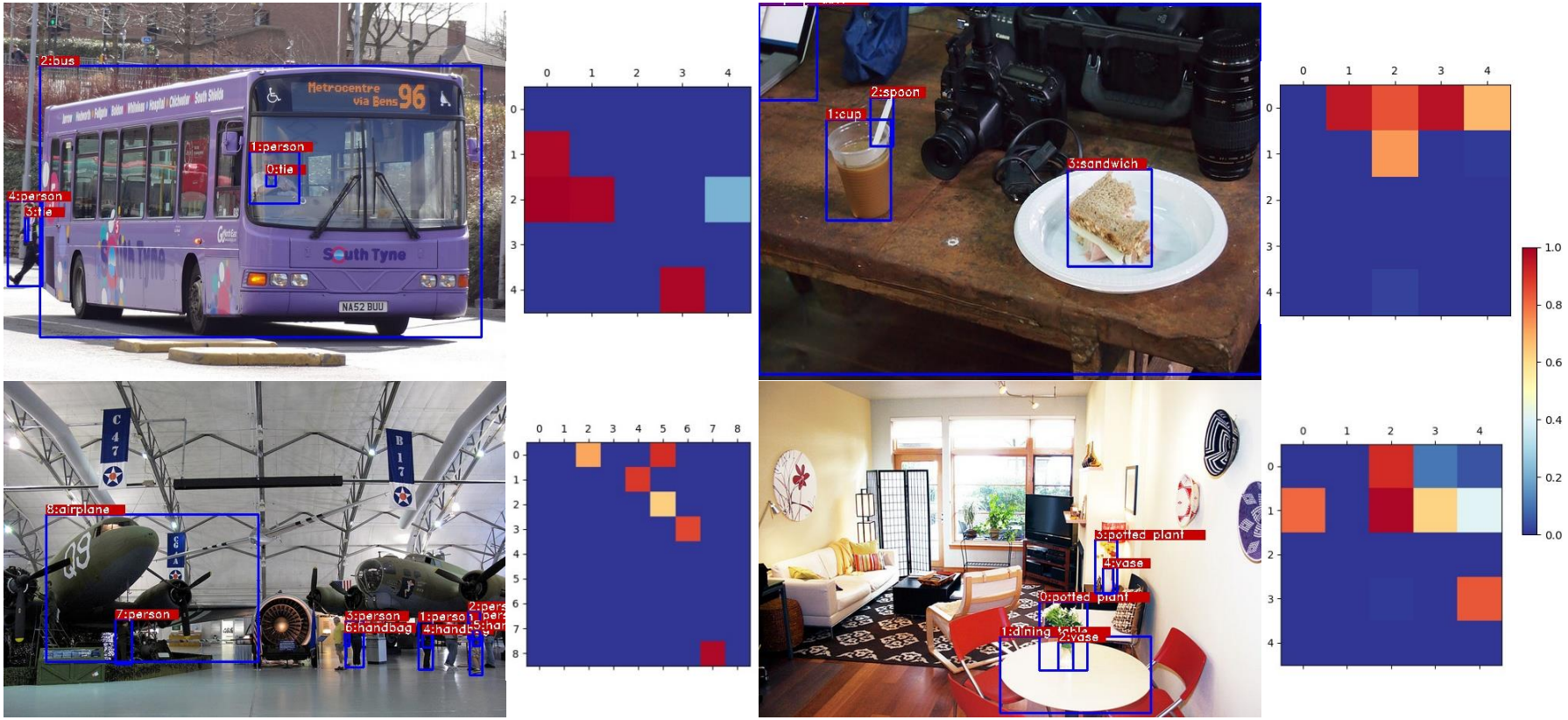
$$\mathcal{L}_R = \frac{1}{N_{inst}^2} \left\| O + O^T - R \right\|_F^2, \tag{12}$$

Ablation Study

Models	PQ	SQ	RQ	PQ^{th}	PQ^{st}
Other Studies					
Panoptic FPN [6]	39.0	-	-	45.9	28.7
AUNet [10]	39.6	-	-	49.1	25.2
OCFusion [8]	41.2	77.1	50.6	49	29
Comparison with UPSNet (use void prediction)					
UPSNet	42.5	78.1	52.5	48.6	33.4
SOGNet (PH1)	43.1	78.6	53.2	49.3	**33.7**
SOGNet (PH2)	**43.5**	**79**	**53.4**	**50.1**	33.6
Comparison with UPSNet (no void prediction)					
UPSNet	42.2	78.3	52.2	48.0	**33.4**
SOGNet (PH 1)	43.0	78.1	53.1	49.3	33.3
SOGNet (PH 2)	**43.7**	**78.7**	**53.5**	**50.6**	33.2

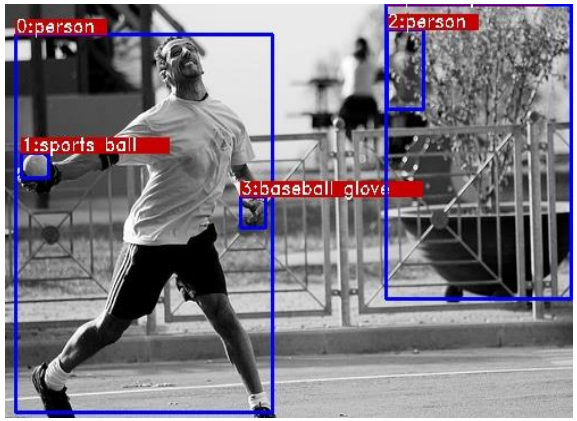
Compare SOGNet with UPSNet and other methods on val set. All models use ResNet-50 as backbone. SOGNet and UPSNet are implemented in the same environment for fair comparison. No augmentation schemes such as the multi-scale training and testing, flipping are used. “PH 1 / 2” denotes the “Panoptic Head 1 / 2”, respectively.

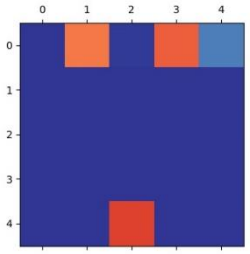
Visual Results

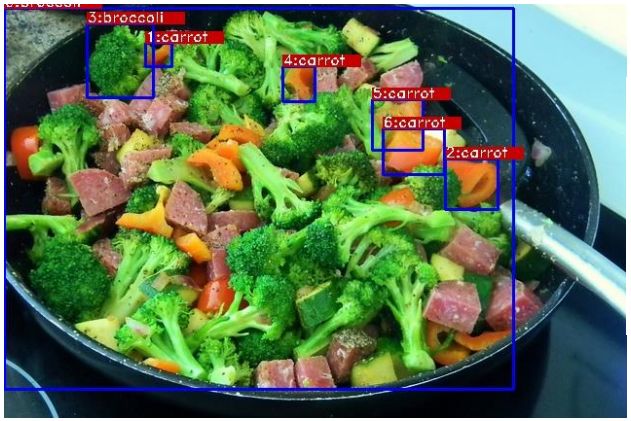


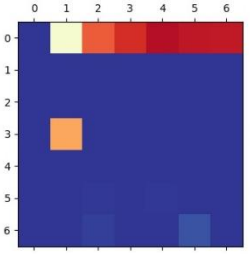
Overlap relations predicted by our method. The map in the right side of each figure is the overlap relation matrix O. Note that the activation on location (I, j) represents that the instance i is covered by (lies below) j.

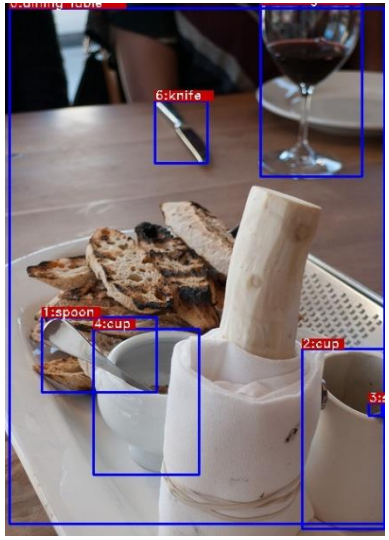
Visual Results

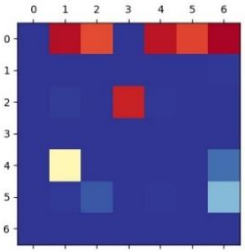


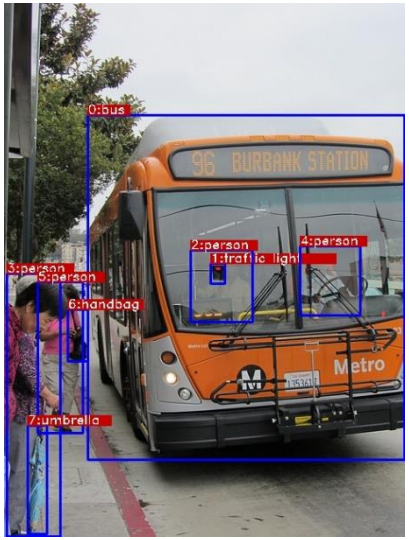


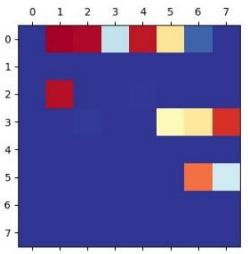


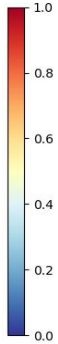






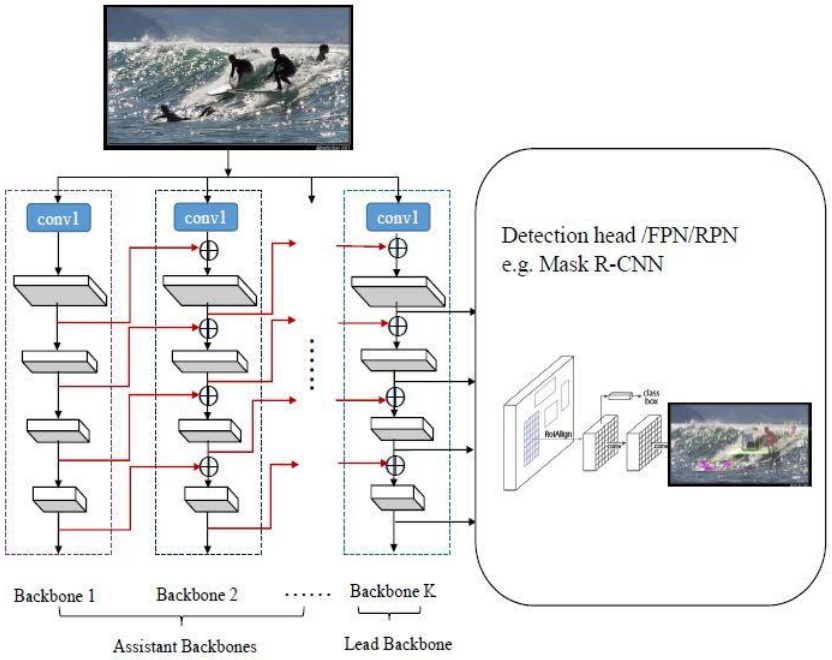






Submitted Entry

CBNet [1]



Method	Backbone	AP_{mask}	AP_{50}	AP_{75}
Cascade Mask R-CNN+ResNeXt152	Single	41	64.1	44.2
	Triple	43.3	66.9	46.8

[1] arXiv preprint arXiv:1909.03625 (2019).

Submitted Entry

Models	backbone	PQ	SQ	RQ
Megvii	ensemble	53.2	83.2	62.9
Caribbean	ensemble	46.8	80.5	57.1
PKU-360	ResNeXt-152	46.3	79.6	56.1
AUNet [10]	ResNeXt-152	46.5	81.0	56.1
UPSNet [16]	ResNet-101	46.6	80.5	56.9
SOGNet	ResNet-101	47.8	80.7	57.6
submitted entry	ResNeXt-152	50.0	81.8	60.0

Submitted entry

Semantic results by SOGNet+ResNet101

+

Instance results by CBNet+ResNeXt (mAP 43.3 on test-dev)

+

Overlap relations by SOGNet+ResNet101

Thank you!

For any question, please contact: ibo@pku.edu.cn