SOGNet: Scene Overlap Graph Network for Panoptic Segmentation

Yibo Yang,[1,2,*] **Hongyang Li,**[2,*] **Xia Li,**[2,3] **Qijie Zhao,**[4] **Jianlong Wu,**[2,5] **Zhouchen Lin**[2,†]
[1]Center for Data Science, Academy for Advanced Interdisciplinary Studies, Peking University
[2]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
[3]Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University
[4]Pony.ai Inc
[5]School of Computer Science and Technology, Shandong University
{ibo, lhy_ustb, ethanlee, zhaoqijie, jlwu1992, zlin}@pku.edu.cn

Abstract

The panoptic segmentation task requires a unified result from semantic and instance segmentation outputs that may contain overlaps. However, current studies widely ignore modeling overlaps. In this study, we aim to model overlap relations among instances and resolve them for panoptic segmentation. Inspired by scene graph representation, we formulate the overlapping problem as a simplified case, named scene overlap graph. We leverage each object's category, geometry and appearance features to perform relational embedding, and output a relation matrix that encodes overlap relations. In order to overcome the lack of supervision, we introduce a differentiable module to resolve the overlap between any pair of instances. The mask logits after removing overlaps are fed into per-pixel instance `id` classification, which leverages the panoptic supervision to assist in the modeling of overlap relations. Besides, we generate an approximate ground truth of overlap relations as the weak supervision, to quantify the accuracy of overlap relations predicted by our method. Experiments on COCO and Cityscapes demonstrate that our method is able to accurately predict overlap relations, and outperform the state-of-the-art performance for panoptic segmentation. Our method also won the Innovation Award in COCO 2019 challenge.

Introduction

Convolutional Neural Networks (CNNs) have achieved huge success in computer vision tasks such as image recognition (He et al. 2016; Yang et al. 2018), semantic segmentation (Long, Shelhamer, and Darrell 2015; Chen et al. 2018), object detection (Girshick 2015; Ren et al. 2015), and instance segmentation (He et al. 2017). The semantic segmentation task answers which background scene a pixel belongs to, while the instance segmentation task predicts foreground object masks. Recently, the panoptic segmentation task introduced in (Kirillov et al. 2019b) aims to unify the results of semantic and instance segmentation into a single pipeline. The system performs semantic segmentation for pixels that belong to amorphous background scenes, named *stuff*. For countable foreground objects, named *things*, the goal is to assign each object region with the right thing class, as well as an instance `id`, identifying which object it belongs to. As a result, panoptic segmentation cannot have overlapping segments. However, most cutting-edge high-performance instance segmentation methods (He et al. 2017) adopt the region-based strategy (Girshick et al. 2014), and output overlapping segments. As shown in Figure 1, the object pairs, such as *cup-dinning table*, *bottle-dinning table*, and *bowl-dinning table*, share overlapping regions from instance segmentation output. Therefore, resolving overlaps and producing coherent segmentation results are the main challenge for the panoptic segmentation task (Kirillov et al. 2019b).

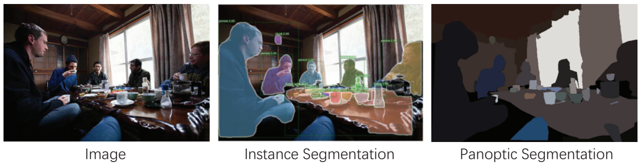

Figure 1: Instance segmentation has overlapping regions for objects, while panoptic segmentation requires a unified result for each pixel. Our study aims to explicitly predict overlap relations and resolve overlaps for the panoptic output.

In (Kirillov et al. 2019b), the semantic and instance segmentation are trained separately, and their panoptic results are merged by heuristic post-processing steps. Later studies aim to unify the semantic and instance segmentation into an end-to-end training framework (Kirillov et al. 2019a; Li et al. 2019b; Liu et al. 2019; Xiong et al. 2019; Porzi et al. 2019; Yang et al. 2019; Li et al. 2018). The panoptic results are usually produced by fusion strategies (Kirillov et al. 2019a; Li et al. 2019b), or predicted by a panoptic head (Liu et al. 2019; Xiong et al. 2019). These studies do not explicitly model overlap relations among objects, which is especially important for datasets with rich categories and complex scenes. However, modeling overlap is challenging without the supervision of object relations or depth information.

In this study, we introduce the scene overlap graph network (SOGNet) for panoptic segmentation. The SOGNet consists of four components: the joint segmentation, the relational embedding module, the overlap resolving module, and the

[*] Equal Contribution

[†] Corresponding Author

panoptic head. The SOGNet trains semantic and instance segmentation in an end-to-end fashion, explicitly encodes overlap relations, resolves the overlap between any pair of objects in a differentiable way, and outputs a unified panoptic result in the panoptic head.

Similar to (Kirillov et al. 2019a; Li et al. 2019b; Liu et al. 2019; Xiong et al. 2019; Porzi et al. 2019; Li et al. 2018), we also use ResNets (He et al. 2016) with feature pyramid network (FPN) (Lin et al. 2017) as the shared backbone for our semantic and instance segmentation branches. Inspired by the relation classification in scene graph parsing tasks (Zellers et al. 2018; Woo et al. 2018), we formulate the overlapping problem in panoptic segmentation as a simplified scene graph with directed edges, in which there are only three relation types for instance i with respect to j: no overlap, covering as a subject, and being covered as an object. We name this representation as *scene overlap graph* in this study. We leverage the category, geometry, and appearance information of objects to perform edge feature embedding for the scene overlap graph, and output a matrix that explicitly encodes overlap relations. However, different from scene graph parsing tasks with the commonly used Visual Genome dataset that has relation annotations, the panoptic segmentation task does not offer annotations of object relations or depth information, so the overlap relations cannot be modeled with direct supervision.

In order to overcome this problem, we develop the overlap resolving module, which resolves the overlaps between any pair of instances in a differentiable way. The mask logits after removing overlaps are then used for per-pixel instance `id` classification in the panoptic head with the panoptic annotation. In doing so, the supervision from pixel-level classification helps the instance-level modeling of overlap relations.

We list the contributions in this study as follows:

- We formulate the overlapping problem in panoptic segmentation as a structured representation, named scene overlap graph. Using category, geometry and appearance features, we perform relational embedding and output a matrix that explicitly encode overlap relations.
- In order to deal with the lack of supervision on overlap relations, we develop an overlap resolving module that resolves overlaps between any pair of instances in a differentiable way. The supervision from per-pixel instance `id` classification in the panoptic head helps to encode overlap relations. We also generate an approximate ground truth as weak supervision to quantify the accuracy of overlap relations predicted by our network.
- Experiments on the COCO and Cityscapes datasets show that, our proposed method is able to accurately predict overlap relations, and outperform the state-of-the-art performance for panoptic segmentation.

Related Work

Image Segmentation The semantic segmentation task focuses on background scenes and is based on fully convolutional networks (FCNs) (Long, Shelhamer, and Darrell 2015). Because detail information is important for dense prediction problems, later studies learn finer representation by deconvolution (Noh, Hong, and Han 2015), encoder-decoder structures (Badrinarayanan, Kendall, and Cipolla 2017), or introducing skip connections between down-sampling and up-sampling paths (Ronneberger, Fischer, and Brox 2015). Other methods aim to aggregate multi-scale context (Farabet et al. 2013; Chen et al. 2018; Zhao et al. 2017), and better capture long-range dependencies (Zheng et al. 2015; Li et al. 2019a). The instance segmentation task deals with foreground objects. Similar to object detection (Girshick 2015; Ren et al. 2015), many instance segmentation studies (Li et al. 2017; He et al. 2017) also adopt the region-based strategy (Girshick et al. 2014), and are able to achieve strong performance due to accurate localization for instances. As another stream, segmentation-based methods (Liang et al. 2018; Arnab and Torr 2017) perform pixel-wise classification and then construct object instances by grouping.

The recently proposed task, panoptic segmentation (Kirillov et al. 2019b), requires a unified result for background scenes and foreground objects. A naive implementation is to train the two sub-tasks separately, and then fuse the results by heuristic rules (Kirillov et al. 2019b). Follow-up studies train semantic and instance segmentation in an end-to-end network by sharing backbone (de Geus, Meletis, and Dubbelman 2018; Kirillov et al. 2019a; Li et al. 2019b; Liu et al. 2019; Xiong et al. 2019; Porzi et al. 2019; Yang et al. 2019; Li et al. 2018). Most of them use fusion heuristics to produce the final output. In (Liu et al. 2019; Xiong et al. 2019), a panoptic head is constructed to predict instance `id`. Li *et al.* (Li et al. 2018) introduce a binary mask to differentiate between thing or stuff for each pixel. A semi- and weakly-supervised method is proposed in (Li, Arnab, and Torr 2018) to relieve the cost of pixel-level annotation.

An important aspect ignored by current panoptic segmentation studies is modeling and resolving overlaps. The study (Lazarow, Lee, and Tu 2019) tries to learn instance occlusions but cannot resolve them in the end-to-end training. As a comparison, our study is able to explicitly model overlap relations, telling us whether an instance lies upon or beneath another, and resolve their overlaps in a differentiable way to generate the panoptic output.

Relational Modeling Parsing relationships of objects has been one of the core components of visual understanding. In (Hu et al. 2018), appearance and geometry features are used to build interactions for object detection. The visual relationship datasets, such as Visual Genome, inspire a series of studies on scene graph generation. In (Zellers et al. 2018; Woo et al. 2018), the low-rank outer product (Kim et al. 2017) is adopted to perform relational embedding from object features. Other relation reasoning methods are proposed by graph-based propagation (Xu et al. 2017), associative embedding (Newell and Deng 2017), and introducing an efficient module (Santoro et al. 2017).

In our study, we formulate the overlapping problem as a simplified scene graph, and also perform relational embedding to encode overlap relations. Our method differs from these studies in that our problem does not offer relation annotation to supervise. We use the supervision from panoptic head to help the modeling of overlap relations.

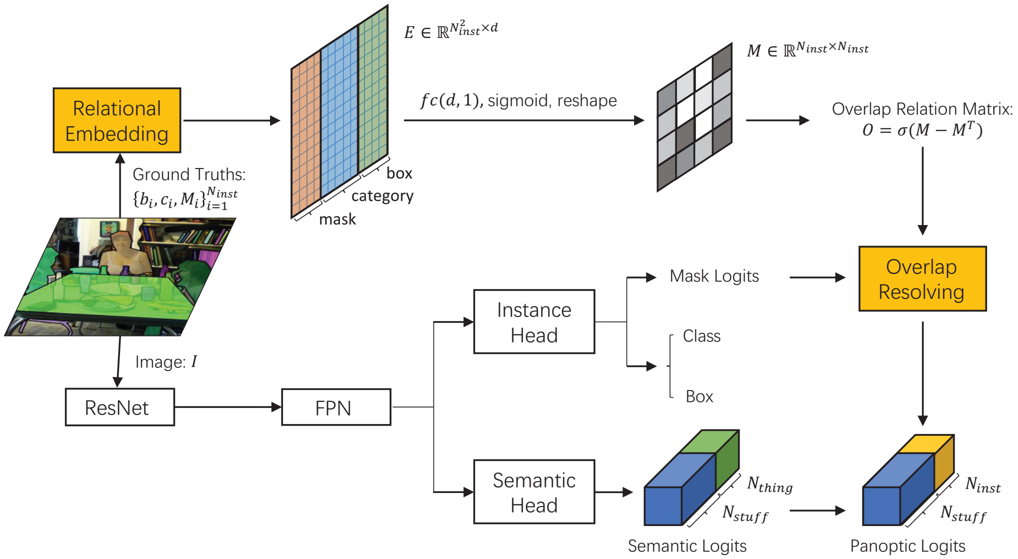

Figure 2: An illustration of the SOGNet for panoptic segmentation. The instance ground truths are input of our relational embedding module. During inference, they are replaced with the predictions from the instance segmentation head. The architecture is trained in an end-to-end manner. σ denotes the ReLU non-linear function.

Scene Overlap Graph Network

In the scene graph generation task (Zellers et al. 2018; Woo et al. 2018; Xu et al. 2017), objects in an image are constructed as a graph and their relations are directed edges. We formulate the overlapping problem in panoptic segmentation as a similar structure, named scene overlap graph (SOG). There are three relation types for instance i with respect to j: no overlap, covering as a subject, and being covered as an object. Our proposed SOGNet consists of four components. The joint segmentation connects semantic and instance segmentation in a unified network. The relational embedding module explicitly encodes overlap relations of objects. After the overlap resolving module, overlaps among instances are removed in a differentiable way. Finally, the panoptic head performs per-pixel instance `id` classification. An illustration of our SOGNet architecture is shown in Figure 2.

Joint Segmentation

Following current popular methods, we use ResNet with FPN as the shared backbone of semantic and instance segmentation branches. The Mask R-CNN structure is adopted for our instance segmentation head, which outputs the box regression, class prediction, and mask segmentation for foreground objects. As for semantic head, the FPN feature maps first go through three 3×3 deformable convolution layers (Dai et al. 2017), and then are up-sampled to the $1/4$ scale. Finally, they are concatenated to generate the per-pixel category prediction. This branch is supervised with both stuff and thing classes, and then the semantic logits of stuff classes are extracted into the panoptic head. We train our model using instance and panoptic annotation. The panoptic annotation that gives per-pixel category and instance `id` supervises the semantic and panoptic head, respectively. The instance annotation contains overlaps and is used for instance segmentation.

Relational Embedding Module

For any training image, we are given the ground truth $\{b_i, c_i, M_i\}_{i=1}^{N_{inst}}$, where b_i, c_i, and M_i refer to the bounding box, one-hot category, and corresponding mask for instance i, respectively, and N_{inst} is the number of instances in this image. As illustrated in Figure 2, we perform relational embedding using the ground truth in the training phase. During inference, we replace them with the prediction from Mask R-CNN branch. The $b_i \in \mathcal{R}^4$ and $c_i \in \mathcal{R}^{80}$ (there are 80 thing classes for COCO) encode geometry and category information, respectively. In order to include appearance feature, we resize the values inside box b_i from M_i as 28×28, which is consistent with the size of Mask R-CNN's output. The resized mask is flattened to be a vector, denoted as $m_i \in \mathcal{R}^{784}$.

The bilinear pooling method learns joint representation for pair of features and is widely applied to visual question answering (Kim et al. 2017; Kim, Jun, and Zhang 2018), and image recognition (Yu et al. 2018) tasks. We construct our category and appearance relation features using the low-rank outer product in (Kim et al. 2017). For a pair of instances i and j, their category relation feature is calculated as:

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

where $\circ$ denotes the Hadamard product (element-wise multiplication), σ is the ReLU non-linear activation, V and U are two linear embeddings that project the input into subject and object features, respectively, and P maps the relation feature into output dimension d_c. We then have the category relation features as:

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

where "[ ]" is the concatenation operation. In a similar way, using m_i as the input of Eq. (1), we can also construct the appearance relation features $E^{(m)} \in \mathcal{R}^{N_{inst}^2 \times d_m}$.

The relative geometry provides strong information to infer whether two objects have overlap or not. Following (Hu et al. 2018; Woo et al. 2018), we have the translation- and scale-invariant relative geometry feature encoded as:

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

where x_i, y_i, w_i, h_i are coordinates and scales extracted from b_i, and $K \in \mathcal{R}^{4 \times d_b}$ is a linear matrix that maps the 4-dimensional relative geometry feature into high-dimensional d_b. We can further have the geometry relation features $E^{(b)} \in \mathcal{R}^{N_{inst}^2 \times d_b}$. We concatenate these edge representations about appearance, category, and geometry as:

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

where $d = d_m + d_c + d_b$. The relational embedding is further used to encode overlap relations.

Overlap Resolving Module

Based on relational embedding, we introduce the overlap resolving module to explicitly model overlap relations and resolve overlaps among instances in a differentiable way.

As illustrated in Figure 2, the relation features, $E \in \mathcal{R}^{N_{inst}^2 \times d}$, go through a $fc(d, 1)$ layer to have a single-channel output with the sigmoid activation to restrict the values within $(0, 1)$. We reshape the output as a square matrix, denoted as $M \in \mathcal{R}^{N_{inst} \times N_{inst}}$. The element M_{ij} has a physical meaning to represent the potential of instance i being covered by instance j. Because there can be only one overlap relation between instances i and j, we then introduce the overlap relation matrix defined as:

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

where σ denotes the ReLU activation that is used to filter out the negative differences between potentials on symmetric positions. In doing so, if $O_{ij} > 0$, it encodes that instance i is being covered by instance j, while on its symmetric position, $O_{ji} = 0$. When $O_{ij} = O_{ji} = 0$, the instances i and j do not have overlaps. Besides, all diagonal elements O_{ii} equals to 0. As explained later, the positive elements in O will be optimized towards 1 in implementations. We now show how to leverage the overlap relation matrix O to resolve overlaps.

For each bounding box, b_i, of the ground truth instances, we have its mask logits (the activations before sigmoid) of 28×28 from the Mask R-CNN output. We then interpolate these logits back to the image scale $H \times W$ by bilinear interpolation and padding outside the box. These logits, denoted as $\{A_i\}_{i=1}^{N_{inst}}$, may have overlaps because Mask R-CNN is region-based and operates on each region independently. Using the matrix O, we can deal with the overlaps between instances i and j as:

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

where A_i' is the output logit of instance i, and s represents the sigmoid activation that turns the logit A_i into a binary-like mask $s(A_i)$. The element-wise multiplication, $s(A_i) \circ s(A_j)$, calculates the intersecting region between instances i and j. The value O_{ij} decides whether the elements in intersecting region should be removed from the logit A_i. When O_{ij} approaches 1, O_{ji} equals to 0, thus the logit A_j will not be affected, and vice versa.

Considering the overlap relations of all the other instances on i, we have:

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

and then the computational steps of the overlap resolving module can be formulated as:

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

where $\mathcal{A} = [A_1, A_2, \cdots, A_{N_{inst}}] \in \mathcal{R}^{H \times W \times N_{inst}}$, and $\times_3$ denotes the Tucker product along the 3-rd dimension (reshape $s(\mathcal{A})$ as $\mathcal{R}^{HW \times N_{inst}}$ for inner product with O^T, and then return to $\mathcal{R}^{H \times W \times N_{inst}}$). We see that our module is friendly to tensor operations in current deep learning frameworks, and is differentiable for resolving overlaps, so that the SOGNet can be trained in an end-to-end fashion.

Panoptic Head

The overlap relation matrix, O, explicitly encodes whether there is intersection between any pair of instances, and if there is, the overlapping region should be removed from which instance. However, we are not provided with the supervision of overlap relations by the panoptic segmentation task. Because accurately resolving overlaps has a strong correlation with the quality of final panoptic output, we can exploit the pixel-level panoptic annotation to assist in the process of modeling overlap relations encoded by O. As illustrated in Figure 2, the instance logits $\mathcal{A}'$ after the SOG module are then fed into the panoptic head.

Following UPSNet (Xiong et al. 2019), we incorporate the logits from semantic head into the mask logits $\mathcal{A}'$. We get the logits of i-th object from semantic output X_i by taking the values inside its ground truth box B_i from the channel corresponding to its ground truth category C_i, and padding zeros outside the box. In UPSNet, they are combined by addition, which is denoted as "Panoptic Head 1". Here we develop an improved combination denoted as "Panoptic Head 2". They are compared as:

$$\text{Panoptic Head 1}: \quad Z_i = X_i + A_i', \tag{9}$$

$$\text{Panoptic Head 2}: \quad Z_i = k \cdot X_i \circ s(A_i') + A_i', \tag{10}$$

where Z_i is the combined logit, s denotes the sigmoid function and k is a factor to balance the numerical difference between semantic output values and mask logits. We set k to be 2 in our experiments. Finally, we concatenate the combined instance logits $\mathcal{Z}_{inst} = [Z_1, ..., Z_{N_{inst}}]$ and the stuff logits $\mathcal{Z}_{stuff}$ from the semantic head to perform per-pixel instance `id` classification with the standard cross entropy loss function, $\mathcal{L}_{panoptic}$.

Despite we do not have the supervision to know which instance lies on the other one, we can leverage the ground truth binary masks, $\{M_i\}_{i=1}^{N_{inst}}$, to infer whether two instances have overlaps or not. We produce a symmetric relation matrix $R \in \mathcal{R}^{N_{inst} \times N_{inst}}$ defined as:

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

where $|\cdot|$ calculates the area of a binary mask through sum operation, $\circ$ calculates the intersection mask through element-wise multiplication, and $\mathbb{1}$ denotes the indicator function that equals to 1 when the condition holds. All diagonal elements R_{ii} are filled with 0. When $R_{ij} = R_{ji} = 1$, it indicates that the overlapped intersection over the smaller object is larger than 0.1, which means there is a significant overlap between instances i and j. With the symmetric relation matrix R, we can introduce the relation loss function as:

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

which calculates the mean squared error between $(O + O^T)$ and R. In doing so, when there is overlap between instances i and j, *i.e.*, $R_{ij} = R_{ji} = 1$, the overlap relation O_{ij} or O_{ji} is forced to approach 1, so that it will not contribute trivially when removing overlaps by Eq. (6).

In total, our SOGNet has the loss functions for semantic and instance segmentation, the panoptic loss $\mathcal{L}_{panoptic}$ for instance `id` classification, and the relation loss $\mathcal{L}_R$ to help optimizing the overlap relation matrix O.

Evaluation Metrics

Panoptic Quality We adopt the evaluation metric introduced in (Kirillov et al. 2019b), called Panoptic Quality (PQ). It can be viewed as the multiplication of a segmentation term (SQ) and a recognition term (RQ):

$$\text{PQ} = \underbrace{\frac{\sum_{(p,q)\in TP} \text{IoU}(p,g)}{|TP|}}_{\text{SQ}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{RQ}}, \tag{13}$$

where p and g are predicted and ground truth segments, and TP, FP and FN denote the true positive, false positive and false negative sets, respectively.

Overlap Accuracy For dataset, such as COCO, the instance annotation permits overlapping instances, while the panoptic annotation contains no overlaps. We can leverage the difference between the two annotations to generate an approximate ground truth of overlap relations, in order to test the quality of overlap relations predicted by our model. The method is also used in (Lazarow, Lee, and Tu 2019) to generate their occlusion ground truth.

Concretely, we are provided with the instance annotation $\{M_i\}_{i=1}^{N_{inst}}$, and the panoptic annotation $\{\hat{M}_i\}_{i=1}^{N_{inst}}$. For any pair of instances i and j, we calculate the intersecting region by $M_i \circ M_j$, and inspect which one of $\hat{M}_i$ and $\hat{M}_j$ mainly covers the intersecting region, to know if i lies upon j or the other way round. Note that the instance and panoptic annotation are not seamlessly matched. Thus this method can only produce approximately true overlap relations

Using the synthetic ground truth as weak supervision, we construct a new asymmetric relation matrix $R^\star$. When $R^\star_{ij} = 1$, we have $R^\star_{ji} = 0$, and it means instance i is covered by j. We can have a new relation loss function in this weakly-supervised setting to replace Eq. (12) with:

$$\mathcal{L}_R^\star = \frac{1}{N_{inst}^2} \left\| O - R^\star \right\|_F^2, \tag{14}$$

which directly supervises the overlap relation matrix O. In experiments, the weakly-supervised manner by Eq. (14) and our method by Eq. (12) have similar performances. Note that the supervision is only valid for datasets such as COCO that has difference between instance and panoptic annotations. It will be ineffective for datasets such as Cityscapes. But our method by Eq.(12) works in both cases.

Thus the weakly-supervised manner by Eq. (14) is served to test the efficacy of our method. Using the weak supervision $R^\star$, we develop a metric, named overlap accuracy (OA), to quantify the quality of overlap predictions encoded by O. The OA of image I is calculated as:

$$\text{OA}(I) = \frac{|TP| + |TN|}{N_{inst} \times (N_{inst} - 1)}, \tag{15}$$

where $TP = \{(i,j) | R^\star_{ij} = 1, O_{ij} \geq 0.5, i \neq j\}$, and $TN = \{(i,j) | R^\star_{ij} = 0, O_{ij} < 0.5, i \neq j\}$. Our reported OA is an average over all images in the validation set.

Experiments

We conduct experiments on the COCO and Cityscapes datasets for panoptic segmentation, and show that our proposed SOGNet is able to accurately predict overlap relations and outperform state-of-the-art performances.

Implementation Details

Training We set the weights of loss functions following (Xiong et al. 2019). The weight of panoptic head is 0.1 for COCO and 0.5 for Cityscapes. The weight of relation loss is set to 1.0. We train the models with a batchsize of 8 images distributed on 8 GPUs. The region proposal network (RPN) is also trained end-to-end. The SGD optimizer with 0.9 Nesterov momentum and a weight decay of 10^{-4} is used. We use an equivalent setting to UPSNet for fair comparison. Images are resized with the shorter edge as 800, and the longer edge less than 1333. We freeze all batch normalization (BN) (Ioffe and Szegedy 2015) layers within the ResNet backbone. For COCO, we train the SOGNet for 180K iterations. The initial learning rate is set to 0.01 and is divided by 10 at the 120K-th and 160K-th iterations. For Cityscapes, we train for 24K iterations and drop the learning rate at the 18K-th iteration. Besides, in order to test the quality of our overlap predictions, we perform an ablation study on COCO using a shorter training schedule because our relation loss converges soon. We only train for 45K iterations and drop the learning rate at iteration 30K and 40K. We do not adopt the void channel prediction proposed in UPSNet. In implementations, we filter out the instances that have no overlap with any other instance to reduce negative samples and computation overhead.

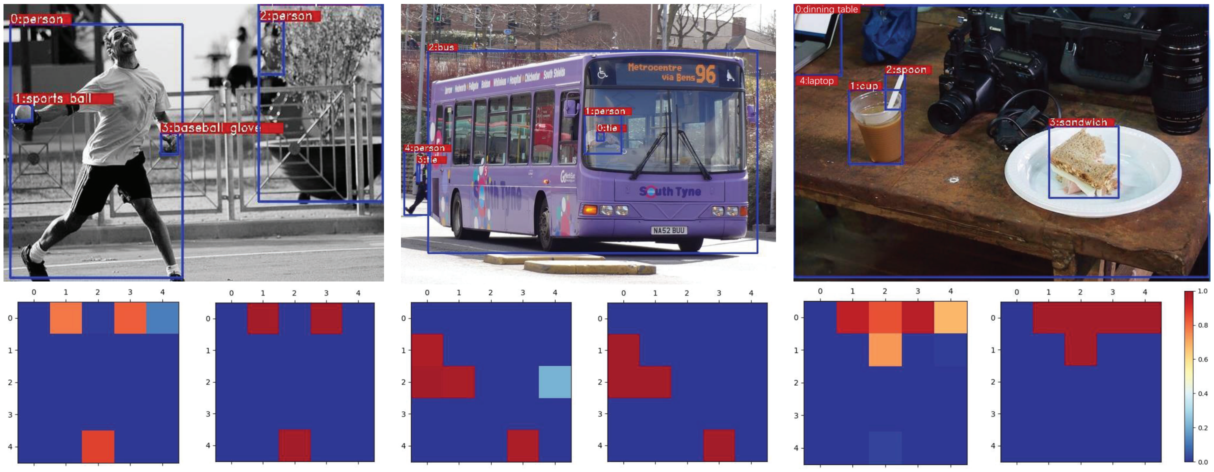

Figure 3: Visualization of the overlap relations encoded by O (down left) and the approximate ground truth, $R^\star$ (down right). Note that the activation on location (i, j) represents that the instance i is covered by (lies below) j. The indices of instances are marked in the images. Zoom in to have a better view. More visualization results can be found in the supplementary material.

Box	Cat	Mask	PQ	SQ	RQ	OA
✓	-	-	37.5	76.3	47.0	69.62
-	✓	-	37.9	76.6	47.3	75.48
✓	✓	-	38.3	76.8	47.7	88.19
✓	✓	✓	**38.4**	**76.9**	**47.8**	**89.22**
Weakly supervised			38.4	77.0	47.7	89.31

Table 1: Different input for the relational embedding module. “Cat”, “Box” and “Mask” denote the category, geometry and appearance features, respectively.

Methods	PQ	SQ	RQ
PlainNet + heuristics	39.6	78.7	48.4
PlainNet + heuristics + label prior	40.9	78.8	49.7
PlainNet + PH1	42.3	78.6	52.1
SOGNet (PH1)	43.0	78.1	53.1
SOGNet (PH2)	**43.7**	78.7	**53.5**

Table 2: PlainNet denotes the joint segmentation component of SOGNet. They are trained in the same condition. “PH 1 / 2” denotes the “Panoptic Head 1 / 2”, respectively.

Inference During inference phase, the ground truths $\{b_i, c_i, M_i\}_{i=1}^{N_{inst}}$ as the input of our relational embedding are replaced with the predictions from Mask R-CNN branch. In order to remove invalid instances, we filter out instances whose probability is lower than a threshold, and perform an NMS-like procedure, following (Kirillov et al. 2019b; Xiong et al. 2019). For highly overlapped predictions of the same class, we keep the mask with the higher confidence score and discard the other one if the intersection is larger than a threshold. Otherwise, we keep the non-interacting part and deal with the next instance. The final output is predicted by our panoptic head. For stuff segment whose area is lower than 4096, we set the corresponding region as void.

Ablation Study

We use ResNet-50 as backbone with a short training schedule, and conduct experiments to analyze feature combinations for our relational embedding, and test the quality of overlap relations predicted by our method. As shown in Table 1, we use different features as the input of our relational embedding. When only category or geometry feature is adopted, the performance improvement on PQ is not so significant, and the overlap prediction does not show high accuracy. When category and geometry features are used together, the embedding becomes much more powerful. Mask feature also slightly improves the overlap accuracy. We expect that a more sophisticated feature design will further boost the performance. It is observed that the weakly-supervised method by Eq. (14) achieves a similar result to our method by Eq. (12). As shown in Figure 3, we visualize the overlap relations predicted by O, as well as the approximate ground truth, $R^\star$, on images from the validation set. More examples can be found in the supplementary material. It is shown that the matrix O accurately predicts some overlap relations, including *baseball glove→person*, *tie→person→bus*, and *spoon→cup→dinning table*. The results demonstrate that the overlap relations are modeled well with the help of supervision from per-pixel instance id classification in the panoptic head. Our method is able to encode overlap relations without direct supervision on them.

Using the standard training schedule and ResNet-50 as the backbone, we also perform comparisons between SOGNet and heuristic inference. The heuristics in (Kirillov et al. 2019b) sort instances according to their objectness scores to deal with overlaps. In (Li et al. 2019b), some hand-crafted label priors are made to rule overlap orders. For example, *tie* should always cover *person*. As a comparison, SOGNet explicitly predict overlap relations and resolve overlaps in a differentiable way. We train the joint segmentation component of SOGNet as a PlainNet, and perform inference with

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-FPN	46.3	79.6	56.1
Panoptic FPN	ResNet-101-FPN	40.9	-	-
OANet	ResNet-101-FPN	41.3	-	-
AUNet	ResNeXt-152-FPN	46.5	81.0	56.1
UPSNet	ResNet-101-FPN*	46.6	80.5	56.9
SOGNet	ResNet-101-FPN*	**47.8**	80.7	**57.6**

Table 3: Comparisons with SOTA performances on COCO *test-dev* set. The first block shows the top-3 entries in public leaderboard of COCO 2018 competition. The second block shows results in recent literatures. $*$ denotes that the backbone has extra deformable convolution layers and longer training schedule is adopted.

different methods. As shown in Table 2, label prior helps to improve the performance. When PlainNet adds the panoptic head for inference to produce the panoptic results, the performance becomes better. The SOGNet with relational embedding and overlap resolving has a further improvement. And our proposed Panoptic Head 2 (PH2) performs better than PH1. In Figure 4, we visualize the panoptic segmentation results of heuristic inference and SOGNet. It is shown that SOGNet better deals with the overlapping problem.

Comparison with Other Methods

We run SOGNet on the COCO and Cityscapes datasets, and compare the results with state-of-the-art methods including the method in (Li, Arnab, and Torr 2018), JSIS (de Geus, Meletis, and Dubbelman 2018), TASCNet (Li et al. 2018), Panoptic FPN (Kirillov et al. 2019a), OANet (Liu et al. 2019), AUNet (Li et al. 2019b), UPSNet (Xiong et al. 2019), and OCFusion (Lazarow, Lee, and Tu 2019).

As shown in Table 3, with ResNet-101-FPN as the backbone, our proposed SOGNet achieves the highest single-model performance on the COCO *test-dev* set. It has a 1.3% PQ improvement than AUNet that uses a larger backbone. SOGNet also performs better than UPSNet using the same backbone and training schedule.

The results of SOGNet on the COCO and Cityscapes validation set are shown in Table 4. It is observed that SOGNet generalizes well to Cityscapes. It has a 0.7% improvement than TASCNet and UPSNet. On the COCO validation set, SOGNet has a 1.2% improvement than UPSNet using the same backbone. The mIoU and AP of SOGNet are 54.56 and 34.2 on COCO, which are similar to the results of UPSNet (54.3 and 34.3 as reported). It indicates that our better panoptic performance is not derived from a stronger semantic or instance segmentation model. More importantly, SOGNet is the only method that can explicitly encode overlap relations and tell us which instance lies upon or beneath another.

Conclusion

In this study, we aim to model overlap relations and resolve overlaps in a differentiable way for panoptic segmentation.

Models	backbone	PQ	PQ^{Th}	PQ^{St}
Cityscapes				
Q.Li *et al.*	ResNet-101	53.8	42.5	62.1
Panoptic FPN	ResNet-101	58.1	52.0	62.5
TASCNet	ResNet-50	59.3	56.3	61.5
UPSNet	ResNet-50	59.3	54.6	62.7
SOGNet	ResNet-50	**60.0**	**56.7**	62.5
COCO				
JSIS	ResNet-50	26.9	29.3	23.3
Panoptic FPN	ResNet-101	40.3	47.5	29.5
OCFusion	ResNet-50	41.2	49.0	29.0
UPSNet	ResNet-50	42.5	48.5	33.4
SOGNet	ResNet-50	**43.7**	**50.6**	33.2

Table 4: Panoptic segmentation results of SOGNet and other state-of-the-art methods on Cityscapes and COCO. Multi-scale testing and flipping are not used.

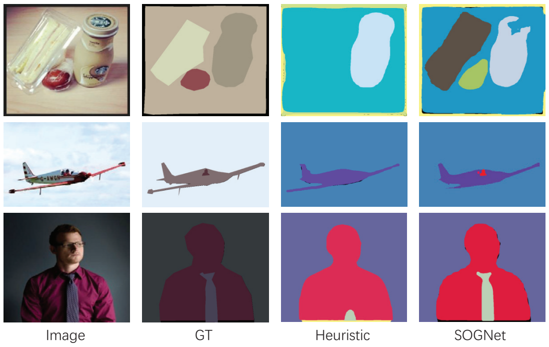

Figure 4: The Visualization of panoptic segmentation results of heuristic inference and SOGNet.

We develop the SOGNet composed of the joint segmentation, the relational embedding module, the overlap resolving module, and the panoptic head. It is able to explicitly encode overlap relations without direct supervision on them. Ablation studies detach SOGNet and analyze the efficacy of each component. Experiments demonstrate that SOGNet accurately predicts overlap relations, and outperforms the state-of-the-art methods on both COCO and Cityscapes.

Acknowledgements

Z. Lin is supported by NSF China (grant no.s 61625301 and 61731018) and Major Scientific Research Project of Zhejiang Lab (grant no.s 2019KB0AC01 and 2019KB0AB02).

References

Arnab, A., and Torr, P. H. 2017. Pixelwise instance segmentation with a dynamically instantiated network. In *CVPR*, 441–450.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI* 39(12):2481–2495.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI* 40(4):834–848.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *ICCV*, 764–773.

de Geus, D.; Meletis, P.; and Dubbelman, G. 2018. Panoptic segmentation with a joint semantic and instance segmentation network. *arXiv preprint arXiv:1809.02110*.

Farabet, C.; Couprie, C.; Najman, L.; and LeCun, Y. 2013. Learning hierarchical features for scene labeling. *TPAMI* 35(8):1915–1929.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 580–587.

Girshick, R. 2015. Fast r-cnn. In *ICCV*, 1440–1448.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*, 2961–2969.

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2018. Relation networks for object detection. In *CVPR*, 3588–3597.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 448–456.

Kim, J.-H.; On, K.-W.; Lim, W.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2017. Hadamard product for low-rank bilinear pooling. In *ICLR*.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. In *NeurIPS*, 1564–1574.

Kirillov, A.; Girshick, R.; He, K.; and Dollár, P. 2019a. Panoptic feature pyramid networks. In *CVPR*.

Kirillov, A.; He, K.; Girshick, R.; Rother, C.; and Dollár, P. 2019b. Panoptic segmentation. In *CVPR*.

Lazarow, J.; Lee, K.; and Tu, Z. 2019. Learning instance occlusion for panoptic segmentation. *arXiv preprint arXiv:1906.05896*.

Li, Q.; Arnab, A.; and Torr, P. H. 2018. Weakly-and semi-supervised panoptic segmentation. In *ECCV*, 102–118.

Li, Y.; Qi, H.; Dai, J.; Ji, X.; and Wei, Y. 2017. Fully convolutional instance-aware semantic segmentation. In *CVPR*.

Li, J.; Raventos, A.; Bhargava, A.; Tagawa, T.; and Gaidon, A. 2018. Learning to fuse things and stuff. *arXiv preprint arXiv:1812.01192*.

Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; and Liu, H. 2019a. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, 9167–9176.

Li, Y.; Chen, X.; Zhu, Z.; Xie, L.; Huang, G.; Du, D.; and Wang, X. 2019b. Attention-guided unified network for panoptic segmentation. In *CVPR*.

Liang, X.; Lin, L.; Wei, Y.; Shen, X.; Yang, J.; and Yan, S. 2018. Proposal-free network for instance-level object segmentation. *TPAMI* 40(12):2978–2991.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*, 2117–2125.

Liu, H.; Peng, C.; Yu, C.; Wang, J.; Liu, X.; Yu, G.; and Jiang, W. 2019. An end-to-end network for panoptic segmentation. In *CVPR*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.

Newell, A., and Deng, J. 2017. Pixels to graphs by associative embedding. In *NeurIPS*, 2171–2180.

Noh, H.; Hong, S.; and Han, B. 2015. Learning deconvolution network for semantic segmentation. In *ICCV*, 1520–1528.

Porzi, L.; Bulò, S. R.; Colovic, A.; and Kontschieder, P. 2019. Seamless scene segmentation. *arXiv preprint arXiv:1905.01220*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 91–99.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. In *NeurIPS*, 4967–4976.

Woo, S.; Kim, D.; Cho, D.; and Kweon, I. S. 2018. Linknet: Relational embedding for scene graph. In *NeurIPS*, 560–570.

Xiong, Y.; Liao, R.; Zhao, H.; Hu, R.; Bai, M.; Yumer, E.; and Urtasun, R. 2019. Upsnet: A unified panoptic segmentation network. In *CVPR*.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *CVPR*.

Yang, Y.; Zhong, Z.; Shen, T.; and Lin, Z. 2018. Convolutional neural networks with alternately updated clique. In *CVPR*, 2413–2422.

Yang, T.-J.; Collins, M. D.; Zhu, Y.; Hwang, J.-J.; Liu, T.; Zhang, X.; Sze, V.; Papandreou, G.; and Chen, L.-C. 2019. Deeperlab: Single-shot image parser. *arXiv preprint arXiv:1902.05093*.

Yu, C.; Zhao, X.; Zheng, Q.; Zhang, P.; and You, X. 2018. Hierarchical bilinear pooling for fine-grained visual recognition. In *ECCV*, 574–589.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *CVPR*, 5831–5840.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *CVPR*, 2881–2890.

Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; and Torr, P. H. 2015. Conditional random fields as recurrent neural networks. In *ICCV*, 1529–1537.