Resolving Fusion Conflicts by Relation Aware Panoptic Network

Lyujie Chen

Tsinghua University, Beijing, China

Abstract. The existing approaches on panoptic segmentation usually learn instance and semantic segmentation separately, thus suffer from fusion conflicts. In this paper, we study the overlapping conflicts in the instance segmentation branch which can be summarized as overlaps between instances and overlaps caused by excessive proposals. We first propose a relation aware module built upon the proposal based instance segmentation model to capture the spatial relation between instances. Then two non-maximum suppression (NMS) like overlapping check strategies are introduced to address the multi-proposal issue during the fusion process. Our **R**elation **A**ware **P**anoptic (**RAP**) segmentation model is generalized to different backbones with consistent accuracy gain up to 2.3% PQ and 3.8% PQ^{Th} on COCO panoptic segmentation benchmark and achieves the state-of-the-art result (49.6 PQ). The source code is available at: `https://github.com/RAPNet/RAP`.

Keywords: panoptic segmentation, fusion conflict, spatial relation

1 Introduction

Panoptic segmentation [14] is a new topic for scene understanding that unifies two distinct tasks of instance segmentation and semantic segmentation. The former detects and segments countable foreground instances (noted *thing*), while the latter focuses on segmenting non-countable background regions (noted *stuff*). Most of the current panoptic segmentation methods [14, 13, 9, 17, 18, 16, 19, 23, 34, 35] deal with this problem by learning two segmentation tasks separately and then fusing the corresponding prediction into a final panoptic output, where each pixel is assigned either to a unique stuff label or specific instance object.

Due to the separation of models, the fusion process will encounter overlapping conflicts, which are not only introduced by the natural inconsistency in the output between instance and semantic segmentation branches but also come from the overlaps between proposals within the instance segmentation branch. The original heuristic fusion method [14] resolves overlaps between instance and semantic segmentation in favor of instances while to simply assign large confidence score objects on top of lower ones. However, the confidence score is not correlated with actual relative spatial position, thus causes obvious occlusion error, as illustrated in Fig. 1.

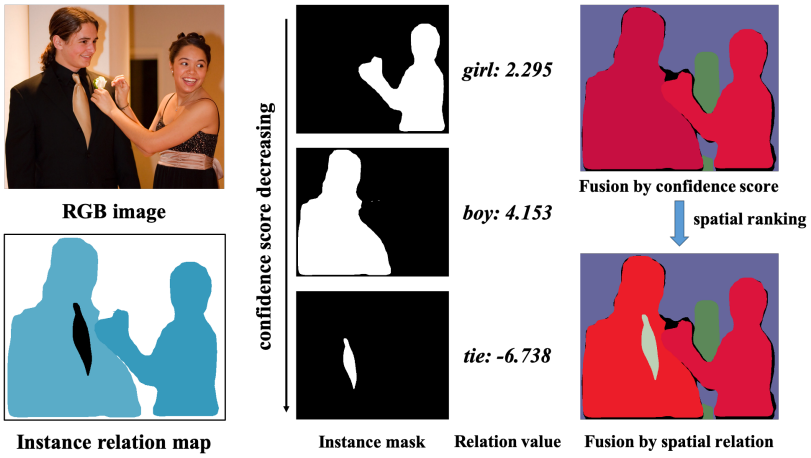

Fig. 1. Our relation aware panoptic segmentation model learns a relation value for each instance. Between overlapped objects, the instance with a smaller relation value (deeper color in the instance relation map) is closer in distance. So we can use the relative spatial relation instead of the detection confidence score to resolve the conflicts.

In this paper, we focus on resolving the overlap conflicts between instances since it may cause more severe occlusion errors, *e.g.*, discarding the entire instance. An intuitive idea is that if we are able to acquire the relative spatial relation between occluded instances, the conflicting pixels can be allocated to the closer object instead of that with a higher confidence score, as illustrated in Fig. 1. Therefore, we propose a lightweight relation aware module added into the existing proposal based instance segmentation network. Supervised by the given overlapping pairs, a spatial relation value for each instance is learned. Between occluded objects, the instance with a smaller relation value is closer in the distance. Based on that, we introduce a new fusion method combining various information, *i.e.*, spatial relation value, instance category, and overlapping area ratio, to further resolve the overlaps caused by excessive proposals. Our **R**elation **A**ware **P**anoptic segmentation model (**RAP**) consistently improves the panoptic quality (PQ) under various network backbones and achieves the state-of-the-art result on COCO panoptic benchmark [22]. The main contributions of this work are as follows:

1. We propose a relation aware module to resolve the overlapping conflicts between instances based on their relative spatial relations.
2. We introduce a fusion method that integrates the learned relative spatial relations with the instance category and overlapping area ratio.
3. We advance the state-of-the-art performance on COCO panoptic benchmark.
4. The relation aware module can be easily applied to the existing proposal based instance segmentation models as an additional head.

2 Related Work

2.1 Panoptic Segmentation

Similar to early works on scene parsing [32] and image parsing [33], panoptic segmentation [14] is a recently proposed multi-task learning problem of instance and semantic segmentation. At present, the divergence of the progress in these two areas brings challenges to joint task learning.

As for instance segmentation, two frameworks are mainly used, namely, segmentation based methods and proposal based methods. The former approaches [3, 24, 15, 1] first classify each pixel and then group them together to obtain the instance objects. The latter methods [29, 7, 20, 10, 25, 12] build upon existing object detection models by performing segmentation within the detected bounding boxes. The proposal based approaches always achieve better performance at the price of introducing overlaps of independent instance predictions in the meanwhile. Semantic segmentation has a rich history in computer vision. With the prevalence of deep learning, FCN [26] based methods have come to the fore. Some early works [2, 31, 28, 30] solve this problem by applying an encoder-decoder architecture to progressively combine high-level and low-level features. Recently, more methods [4, 5, 38, 37] rely on the use of dilated convolution [36] to enlarge receptive field and boost the performance.

Panoptic segmentation, first proposed by Kirillov *et al.* [14], encompasses both *stuff* and *thing* classes to realize a high-level understanding of images. In [14], two individual models are used to learn semantic and instance segmentation separately and the outputs are then merged in a heuristic way. This training strategy doesn't perform any shared computation, thus it is computationally inefficient and cannot fully exploit useful mutual information between two tasks. For this reason, some end-to-end learning methods are proposed. JSIS [9] uses a shared feature extractor backbone for both semantic and instance head. Panoptic Feature Pyramid Network (FPN) [13] endows Mask-RCNN [10] with a lightweight semantic segmentation branch using a shared FPN [21] backbone. AUNet [19] further adds two sources of attention to a unified framework to use complementary information for better background segmentation. However, although learning in an end-to-end way, the conflicts of outputs from two branches are still inevitable.

2.2 Fusion Methods

The original fusion method in [14] comes up with the following rules to resolve overlapping problems: (1) retain instance objects with higher confidence score when overlap occurs, (2) resolve overlaps between *stuff* and *thing* in favor of *thing*, and (3) discard the entire stuff segment labeled *other* or under a given area threshold. In addition, to handle the issue of excessive proposals, the instance will be discarded in two situations: (1) instance with confidence score lower than confidence threshold ρ, or (2) overlapping area ratio between instance and existing assigned region is higher than overlap threshold τ (usually 0.5). However,

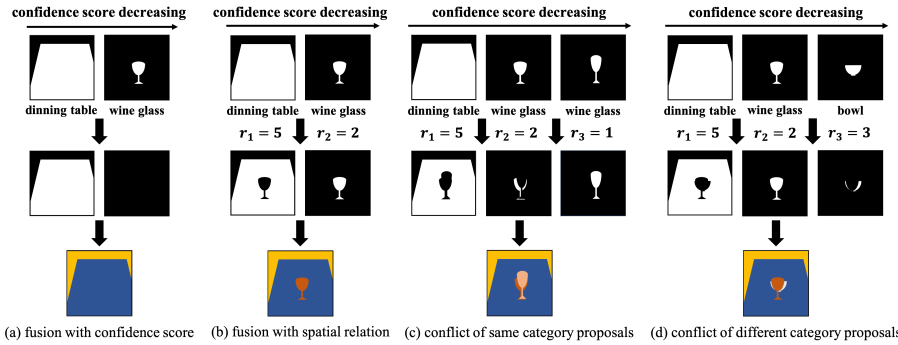

Fig. 2. Illustration of fusion conflicts between instances. Given an image with a wine glass on a dining table, (a)(b) consider the situation that each object has one proposal, while (c)(d) handle the situation with multiple proposals. In each sub-image, the original mask and its relation value (if exists) of each proposal predicted by the model are presented in the first row. The second row is refined masks based on different fusion strategies and the third row is the corresponding panoptic result.

in most cases, the objects with a larger size or in more common categories tend to achieve a higher detection confidence score. So this fusion strategy causes severe occlusion errors. Take Fig. 2 (a) as an example, a dining table with a higher score is allocated before a wine glass, so the glass will be discarded because of the overlapping area ratio exceeding the threshold. As a result, it is not desirable to regard the confidence score as the standard to keep instances. In our approach, we instead use the instance spatial relation as well as comprehensively take the object category and overlapping area ratio into consideration for fusing the final result.

Recently, a series of methods [23, 17, 34, 16] have been proposed to solve this problem. OANet [23] introduces a spatial ranking module to deal with simple occlusion errors by learning spatial relations at the category-level. But this method fails to handle the overlaps between the same category instances or in some complex situations, such as people can either stand behind or sit inside the bus. TASCNet [17] puts forward things and stuff consistency (TASC) network to decide whether to assign pixels to *thing* or *stuff*, which resolves the conflict between instance and semantic segmentation to some extent. UPSNet [34] proposes a parameter-free panoptic head to tackle two types of conflicts at the same time via pixel-wise classification. However, the improvement is limited (about 0.4% PQ improvement) due to its oversimplified design and lack of interpretability. OCFusion [16] is a recent similar work to our idea, which also learns spatial occlusion at the instance-level. They use the learned relations to help make decisions during the fusion process, which partly solves the conflict. In comparison, we specifically analyze the reason for instance conflicts and come up with an elaborated fusion strategy, achieving more improvement in panoptic quality and better time efficiency at inference.

3 Method

As described in Section 2.2, the heuristic fusion method [14] resolves overlaps between different instances by preferentially assigning pixels to instances with a higher confidence score. However, this score is not correlated with how well instance masks are segmented [12] or whether spatial relations are correct, thus resulting in a variety of occlusion errors. In this section, we address this issue by proposing a relation aware module to learn instance-level spatial relations. Specifically, we introduce the generation method for ground truth of spatial relations, the network structure of relation aware module as well as its training strategy. We further explain the meaning of the learned relation value and discuss the advantages of our learning paradigm. Then, we detail two situations that overlap conflicts occur within the instance segmentation branch and introduce our fusion method which incorporates various information to solve them.

3.1 Ground Truth Spatial Relations

In order to predict spatial relations, the first step is to extract the ground truth data. We extract the spatial relations of occluded instances by comparing the difference of ground truth segmentation maps between instance and panoptic segmentation tasks. Specifically, given two masks from instance segmentation I_i and I_j. Their overlapping region is denoted as $O_{ij} = I_i \bigcap I_j$. We define the overlapping area ratio as: $R_i = S(O_{ij})/S(I_i)$ and $R_j = S(O_{ij})/S(I_j)$, where $S(I)$ represents the pixel number of the mask I. If either R_i or R_j exceeds the threshold θ, it can be speculated that I_i and I_j significantly overlap each other. If the majority of the pixels in the overlapping area are assigned to I_i in panoptic segmentation ground truth map (due to the inaccurate labeling, overlapping pixels might be assigned to more than one instance), we could extract the spatial relation that I_i is in front of I_j, denoted as $F_{ij} = 1$ or $F_{ji} = -1$.

3.2 Learning with Relations

Our learning strategy is inspired by [6] where pixel-level ordinal relation pairs are used for depth prediction problems. We extend this idea and learn the per-instance relation values in a weakly supervised way. In detail, we use a ranking loss to supervise the relative consistency between different instances. Given the extracted pairwise spatial relation F_{ij} of I_i and I_j, and their corresponding predicted relation value v_i and v_j, the relation loss L_r is defined as:

$$L_r(I_i, I_j) = \begin{cases} log\left(1 + exp\left(v_i - v_j\right)\right), & F_{ij} = 1 \\ log\left(1 + exp\left(-v_i + v_j\right)\right), & F_{ij} = -1 \end{cases} \tag{1}$$

$$L_r = \frac{1}{K} \sum_{(I_i, I_j, Fij)}^{\mathcal{F}} L_r(I_i, I_j) \tag{2}$$

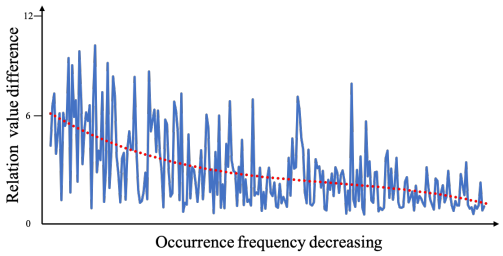

Fig. 3. The average numerical difference of relation value on different overlapping types in the descending order of their occurrence frequency.

where $\mathcal{F}$ is the collection of all known relation pairs with a size of K. It is easy to see from the formula that this ranking loss encourages a large difference between relation values. In this way, the predicted relation value v of each instance is implicitly learned and updated.

What does relation value mean? (1) According to the above formula, the instance that closer in distance will be learned to have a smaller relation value. Although it looks similar to the concept of *depth* or *ordinal depth*, it is actually different. Relation value is an unbounded real number, thus cannot represent the absolute distance like *depth*. In addition, since only overlapped instance pairs are supervised during training, the learned instance-level relation value isn't globally consistent and transitive, which results in (i) the comparison of the relation value is meaningful just between two **occluded** instances and (ii) only the **numerical difference** of the relation value is meaningful. Therefore, this value cannot be simply interpreted as *ordinal depth* as well. It should be used to indicate the relative spatial position between occluded instances to resolve the conflicts. (2) As shown in Fig. 3, we have calculated the average difference of the relation value of each overlapping type (*e.g.*, *person* and *tie*, *table* and *bowl*) according to the prediction result on the COCO [22] val dataset. The numerical difference is sorted in descending order of the occurrence frequency of overlap type in the COCO train dataset. The red trendline fitted with a polynomial of order 3 shows a downward trend, indicating that the relation value difference between more common overlapping types of instance pairs is generally larger. Also, since commonly appeared overlaps are always learned better, we infer that the **difference of the relation value can reflect the confidence of the prediction**. For example, in Fig. 1, the relation value difference between *tie* and *boy* is significantly greater than that between *boy* and *girl*. This is because the former situation is more common and easier to predict, while the latter one involved the same category instances is more difficult.

Advantages Compared to OCFusion [16] which formulates the spatial relation learning as a classification problem, our regression learning paradigm seems to be less straightforward. However, it solves two significant issues, *i.e.*, classification ambiguity and poor time efficiency: (1) Since OCFusion predicts the relative position of each instance pair independently, it may appear contradictory, *e.g.*,

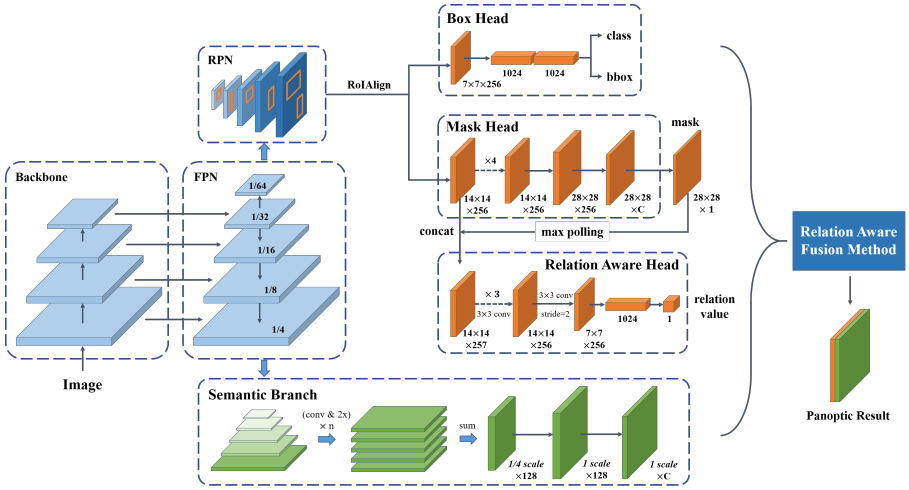

Fig. 4. Network architecture. We build relation aware panoptic network by adding a relation aware head based on panoptic FPN [13].

A in front of B, B in front of C, but C in front of A. However, the relation value given by our model is a unique real number, so the overlapped instances have a certain relative spatial position. (2) More importantly, during the test phase, our learning strategy is time-efficient since the time complexity for inference of relation value is a stable linear function $O(N)$ regardless of the occurrence number of overlaps. In contrast, OCFusion learns the occlusion between two instances and costs $O(N^2)$ times of inference during the fusion process in the worst case. In Section 4.3, we will further compare the effectiveness and efficiency of these two methods in detail.

3.3 Network Architecture

Since the spatial relation value is predicted for each instance, it can be regarded as another instance attribute besides class, bounding box and mask. So an intuitive idea is to add a "head" to the proposal based instance segmentation model to regress this value. In this paper, we use panoptic FPN [13] as the basic panoptic segmentation model. Based on that, we build a relation aware head upon Mask R-CNN referring to the architecture in [12]. As shown in Fig. 4, its input is the combined features in mask head and the relation value is predicted through four layers of 3×3 convolutions and one fully connected layer. According to [13], the total loss can be extended as

$$L = \lambda_i(L_c + L_b + L_m) + \lambda_s L_s + \lambda_r L_r \tag{3}$$

where L_c, L_b, L_m and L_s are the classification loss, bounding-box loss, mask loss and semantic segmentation loss respectively. The hyperparameters λ_i, λ_s and λ_r are used to balance the loss.

Algorithm 1 Fusion strategy between instances

P and V are $H \times W$ matrices, initially empty and infinite respectively. ρ, γ, τ are the hyperparameters. s_i, r_i, l_i are the corresponding confidence score, relation value and label for proposal I_i respectively.

sort $\{I_i\}$ in descending order of s_i.
discard I_i with $s_i < \rho$
for each proposal I_i **do**
    $O_i = I_i \cap (P == l_i)$
    $L_i = S(O_i)/S(I_i)$
    if $L_i \geqslant \tau$ **then**
        continue
    end if
    for $j < i$ **do**
        $O_{ij} = I_i \cap I_j$
        $R_i = S(O_{ij})/S(I_i)$
        $R_j = S(O_{ij})/S(I_j)$
        if $R_i \geqslant \gamma$ and $R_j \geqslant \gamma$ **then**
            discard proposal I_i
        end if
    end for
    $C_i = I_i \cap (V > r_i)$
    if $S(C_i)/S(I_i) \leqslant \tau$ **then**
        continue
    else
        assign l_i to P at mask C_i
        assign r_i to V at mask C_i
    end if
end for

3.4 Fusion Strategy

In order to handle the overlapping pixels during the fusion process, we analyze two types of instance conflicts in the proposal based instance segmentation, (1) overlaps between different instances and (2) overlaps caused by excessive proposals, from which our final fusion strategy is decided.

The first type of conflict always leads to abandon of meaningful instances. As illustrated in Fig. 2 (a), by using the fusion strategy described in Section 2.2, the table will be put on the canvas first and the wine glass is completely discarded due to the excessive large overlapping area ratio. This mistake can be easily addressed with the help of instance spatial relations. Before calculating the overlap area ratio, the 'real' mask of the proposal can be refined by assigning the overlapping pixels to the instance with a smaller relation value (*i.e.*, closer in distance). In this way, due to the awareness that the wine glass is put on top of the dining table, the mask of the table will be correctly refined with no overlap as shown in Fig. 2 (b).

The second type of conflict generally retains too much incorrect instances since an object might be predicted with multiple proposals either of the same cat-

egory (as shown in Fig. 2 (c)) or of the different category (as shown in Fig. 2 (d)). The former mistake can be addressed by extending the check of the overlapping area ratio to a category-aware way (named *category-aware check*). The proposal will be discarded only if its overlapping area ratio to the already assigned region of the same class exceeds the threshold τ. To tackle the problem in Fig. 2 (d), we propose an *excessive overlap check* which calculates the overlapping area O_{ij} between the new proposal I_i with every assigned proposal I_j. If the overlapping ratio R_i and R_j *both* exceed the threshold γ, the new proposal will be discarded.

In summary, our final fusion strategy between instances can be described in Algorithm 1. After that, the fusion process between instance and semantic segmentation branches remains the same as [14]. Finally, segments that labeled other or under the given area threshold (4096 for COCO) are discarded.

4 Experiment

4.1 Implementation Details

COCO Dataset We perform experiments on COCO 2019 panoptic segmentation dataset [22] which consists of 118K images for training, 5K images for validation and 20K images for test-dev. The corresponding annotations consist of 80 *thing* categories and 53 *stuff* categories. We train our model on the training set and evaluate on the validation set. Besides, we submit results on test-dev to COCO 2019 panoptic segmentation leaderboard.

Metrics We adopt the panoptic quality (PQ) metrics introduced in [14]. It can be regarded as the product of segmentation quality (SQ) and recognition quality (RQ). From another perspective, PQ can be decomposed into scores specific to things and stuff, denoted as $\mathrm{PQ}^{\mathrm{Th}}$ and $\mathrm{PQ}^{\mathrm{St}}$, respectively.

Implementation Based on Mask R-CNN benchmark [27], we implement the panoptic FPN [13] and our RAP network by adding a semantic segmentation branch and a relation aware head described in Section 3.3. Different from the design in [13], the deepest FPN level in the semantic segmentation branch is at 1/64 scale. The source code is available here: `https://github.com/RAPNet/RAP`.

Firstly, we train the panoptic FPN [13] without relation aware head for 150K iterations on 8 NVIDIA P40 GPUs with a batch size of 16, where $\lambda_i = 1.0$ and $\lambda_s = 0.5$. The initial learning rate of 0.02 is reduced by 10 at 80K and 120K iterations. We use the SGD as the optimization algorithm with momentum 0.9 and weight decay 0.0001. Here we spend more time on training than original Mask R-CNN does (90K iterations) since we find that the semantic branch needs a longer time to converge. In this way, about 0.4∼0.9 PQ improvement is obtained on different backbones (more effective for shallower backbone). After that, we train relation aware head for 5K more iterations by fine-tuning with all other parameters frozen, where $\lambda_r = 1.0$. The learning rate is 0.002 at the beginning and reduced by 10 at 2.5K iterations. We set threshold $\theta = 0.2$ for extracting the spatial relations ground truth in COCO.

Table 1. Sensitivity analysis of various threshold hyperparameters ρ, τ, γ with the backbone of ResNet-101.

	$\tau = 0.5$			$\tau = 0.6$		
	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.7$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.7$
$\rho = 0.5$	+1.579 PQ +2.624 PQ^{Th}	+1.559 PQ +2.590 PQ^{Th}	+1.533 PQ +2.548 PQ^{Th}	+1.627 PQ +2.678 PQ^{Th}	+1.598 PQ +2.631 PQ^{Th}	+1.577 PQ +2.594 PQ^{Th}
$\rho = 0.6$	+1.906 PQ +3.102 PQ^{Th}	+1.894 PQ +3.083 PQ^{Th}	+1.881 PQ +3.062 PQ^{Th}	**+1.941** PQ **+3.144** PQ^{Th}	+1.921 PQ +3.112 PQ^{Th}	+1.908 PQ +3.090 PQ^{Th}

Table 2. Ablation study on fusion strategies. Relation aware module, category-aware check, and excessive overlap check are shorted as RA module, CA check and EO check respectively. * is our implementation which uses more features in the semantic branch and trains a longer time, same as RAP.

Backbone	Method				PQ	SQ	RQ	PQ^{Th}	PQ^{St}
ResNet-50	Panoptic FPN [13]				39.0	-	-	45.9	28.7
	Panoptic FPN*				40.0	78.0	49.1	46.2	30.5
	Ours	RA module	CA check	EO check					
		√			40.8	77.9	50.3	47.6	30.5
		√	√		41.5	78.1	51.0	48.8	30.5
		√	√	√	41.8	78.1	51.3	49.2	30.5
	Relative improvement				**+1.8**	**+0.1**	**+2.2**	**+3.0**	**+0.0**
ResNet-101	Panoptic FPN [13]				40.3	-	-	47.5	29.5
	Panoptic FPN*				41.4	79.5	50.7	47.8	31.7
	Ours	RA module	CA check	EO check					
		√			42.6	79.6	52.2	49.7	31.8
		√	√		43.0	79.6	52.7	50.5	31.8
		√	√	√	43.3	79.6	53.0	50.9	31.8
	Relative improvement				**+1.9**	**+0.1**	**+2.3**	**+3.1**	**+0.1**

4.2 Ablation Study

Hyperparameters Analysis During the inference, the threshold hyperparameters ρ, τ, γ is closely related to the final result. Therefore, we first perform a sensitivity analysis on these three thresholds. We use $\rho = 0.5$ and $\tau = 0.5$ for panoptic FPN [13] with the heuristic fusion method (our implementation) as the baseline and report the relative improvement of RAP with different hyperparameters choice in Table 1. The result indicates that ρ is more sensitive than τ and γ on different values and the best result is achieved when $\rho = 0.6$, $\tau = 0.6$ and $\gamma = 0.5$. So in the following experiments, we adopt this hyperparameter choice for our RAP network.

Fusion Strategies We next analyze the contribution of our *relation aware module*, *category-aware check* and *excessive overlap check* on two different backbones. As shown in Table 2, these three strategies are all beneficial to the final result while the overall relative improvement on PQ and PQ^{Th} are up to 1.9 and 3.1 point. It demonstrates that too many meaningful instances will be discarded

Table 3. Evaluation of RAP with different backbones on COCO val dataset. The value in the bracket is the relative improvement compared to panoptic FPN (our implementation) [13] trained with the same settings. 'dc' means deformable convolution.

Backbone	PQ	RQ	PQ^{Th}
ResNet-50	41.8 (+1.8)	51.3 (+2.2)	49.2 (+3.0)
ResNet-101	43.3 (+1.9)	53.0 (+2.3)	50.9 (+3.1)
ResNeXt-101	45.7 (+2.3)	55.8 (+2.6)	53.9 (+3.5)
ResNeXt-152	46.9 (+2.3)	57.0 (+2.6)	55.5 (+3.6)
ResNeXt-152 (w/ dc)	49.4 (+2.4)	59.6 (+2.7)	57.2 (+3.8)

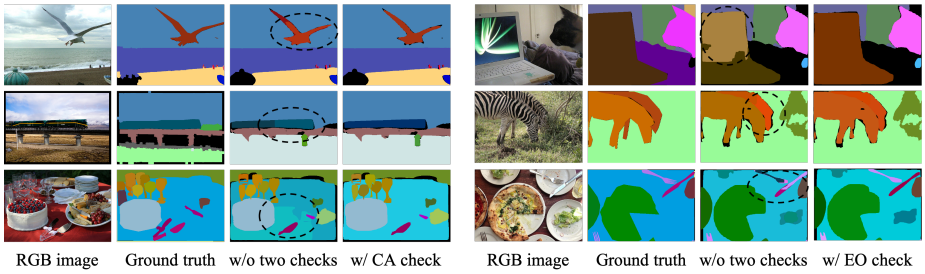

Fig. 5. Illustration of two overlapping area ratio check strategies. *Category-aware check* and *excessive overlap check* are shorted as CA check and EO check respectively. We point out the redundant proposals with the black circle. By using the proposed strategies, they are all correctly discarded.

if simply using the overlap ratio to get rid of redundant proposals as heuristic fusion methods do. What's more, as illustrated by both metrics in Table 2 and visualization result in Fig. 5, it is observed that the relation aware module cannot handle the situation that one instance has multiple proposals. As a result, the combination of these three strategies is crucial to address these two types of conflicts at the same time and yields a stable improvement.

Different Backbones We generalize our method to different network backbones as shown in Table 3. In all settings, RAP obtains consistent panoptic quality gain. We also find that with stronger backbone used, the more relative improvement of our method is achieved.

4.3 Comparison to OCFusion

OCFusion [16] is a recent similar work that resolves the instance conflicts by learning occlusion between two objects as a classification problem. Its fusion strategy only considers the instance relation which is very similar to not using overlapping area ratio check strategies in our method. To better compare the differences, we investigate the prediction accuracy of the instance spatial relations and the inference time efficiency. Due to no source code provided, we implement OCFusion using the same backbone and learning strategy as RAP.

Table 4. Comparison to OCFusion on COCO val dataset with the backbone of ResNet-101. * is our implementation under exactly the same settings.

Method	Acc.	PQ	RQ	PQ^{Th}
Panoptic FPN [13]*	36.56%	41.4	50.7	47.8
OCFusion [16]*	96.98%	42.5	52.2	49.6
RAP	96.31%	43.3	53.0	50.9

Table 5. Comparison to OCFusion of inference time efficiency with the backbone of ResNet-101. Inference time on occlusion head of OCFusion includes the time for the construction of input feature matrix which is not needed in our pipeline. * is our implementation under exactly the same settings.

Method	Relation inference	Network inference	Fusion process
OCFusion [16]*	136.7ms	274.3ms	22.3ms
RAP	0.7ms	137.8ms	19.6ms
Acceleration	**195.3**×	**1.99**×	**1.14**×

Without knowing all the details of its original implementation, it's inevitable to have some numerical differences from its original paper. However, except for the different spatial relations learning strategies and the final fusion process, we keep all the settings the same between RAP and OCFusion. So we believe this comparison is very fair and reasonable.

Relation Prediction Accuracy Within the loop of lines 9 to 16 in Algorithm 1, we calculate the relation prediction accuracy between the new proposal and allocated proposals with known spatial relations. As we can see from Table 4, although the accuracy of RAP is slightly lower than OCFusion, there is an obvious gap in the improvement of panoptic quality. It mainly benefits from the use of two NMS like overlapping ratio check strategies during the fusion process which deals with the multi proposal conflicts.

Inference Efficiency During the test phase, the fusion process between OCFusion and RAP are very similar, so the time difference mainly comes from the model inference stage. In theory, since OCFusion predicts the occlusion relations of every pair of instances. it will operate network inference under $O(n^2)$ time complexity in the worst case. On the contrary, our method only needs consistently $O(n)$ times of inference. In practice, for an image with n proposals, our approach only performs one inference with batch size of n on GPU. But as for OCFusion, it can either perform m times of inference with batch size of 1 when overlap occurs or predict all n^2 occlusion relations by one inference with a batch size of n^2. m is the number of significant overlap occurrences. The first strategy is time-consuming since it cannot take advantage of GPU parallel computing. The second strategy must take some time to construct the input matrix of the occlusion head and also suffers from the insufficient GPU memory when n is too large (*e.g.*, $n > 50$ for NVIDIA P40 GPU).

Table 6. Comparison to prior works on COCO val dataset with ResNet-50 and ResNet-101 backbones.

Method	Backbone	PQ	SQ	RQ	PQ^{Th}	PQ^{St}
Panoptic FPN [13]	ResNet-50	39.0	-	-	45.9	28.7
Panoptic FPN [13]	ResNet-101	40.3	-	-	47.5	29.5
OANet [23]	ResNet-50	39.0	77.1	47.8	48.3	24.9
OANet [23]	ResNet-101	40.7	78.2	49.6	50.0	26.6
AUNet [19]	ResNet-50	39.6	-	-	49.1	25.2
UPSNet [34]	ResNet-50	42.5	78.0	52.4	48.5	33.4
OCFusion [16]	ResNet-50	41.3	-	-	49.4	29.0
OCFusion [16]	ResNet-101	43.0	-	-	51.1	30.7
RAP	ResNet-50	41.8	78.1	51.3	49.2	30.5
RAP	ResNet-101	43.3	79.6	53.0	50.9	31.8

Table 7. Comparison to the state-of-the-art methods on COCO test-dev dataset. The value in the bracket is the relative improvement compared to panoptic FPN (our implementation) [13] trained with the same strategies and backbones. 'dc' means deformable convolution. † use multi-scale testing.

Method	Backbone	PQ	SQ	RQ	PQ^{Th}	PQ^{St}
Megvii (Face++)	ensemble	53.2	83.2	62.9	62.2	39.5
Caribbean	ensemble	46.8	80.5	57.1	54.3	35.5
PKU 360	ResNeXt-152	46.3	79.6	56.1	58.6	27.6
Panoptic FPN [13]	ResNet-101	40.9	-	-	48.3	29.7
OANet [23]	ResNet-101	41.3	-	-	50.4	27.7
AUNet [19]†	ResNeXt-152	46.5	81.0	56.1	55.8	32.5
UPSNet [34]†	ResNet-101 (w/ dc)	46.6	80.5	56.9	53.2	36.7
OCFusion [16]†	ResNeXt-101 (w/ dc)	46.7	-	-	54.0	35.7
RAP	ResNeXt-152	**47.0 (+1.9)**	**80.4 (+0.3)**	**57.1 (+2.3)**	**55.8 (+3.1)**	**33.7 (+0.1)**
	ResNeXt-152 (w/ dc)	**49.6 (+2.3)**	**81.4 (+0.2)**	**59.9 (+2.6)**	**57.8 (+3.8)**	**37.4 (+0.1)**

Here, we compare our method to OCFusion using the second strategy. The experiment is conducted on a NVIDIA P40 GPU with 24GB memory. As illustrated in Table 5, our method takes much less time than OCFusion on model inference. Also, since all the spatial relations are certain and can be obtained at one time, RAP deals with the allocation of overlapping pixels faster, resulting in less fusion process time.

4.4 Comparison to Existing Methods

In the end, we compare RAP with the state-of-the-art methods on COCO [22] val and test-dev set. We adopt ResNet-50-FPN as well as ResNet-101-FPN backbones [11] for val dataset and ResNeXt-152-FPN (w/ and w/o deformable convolution [8]) backbone for test-dev set. It is observed from Table 6 and Table 7 that RAP achieves very competitive or best performance compared to existing approaches, even in the comparison with those multi-scale testing methods (UPSNet [34], AUNet [19], and OCFusion [16]). As shown in Fig. 6 and Fig. 7, the occlusion errors can be well addressed.

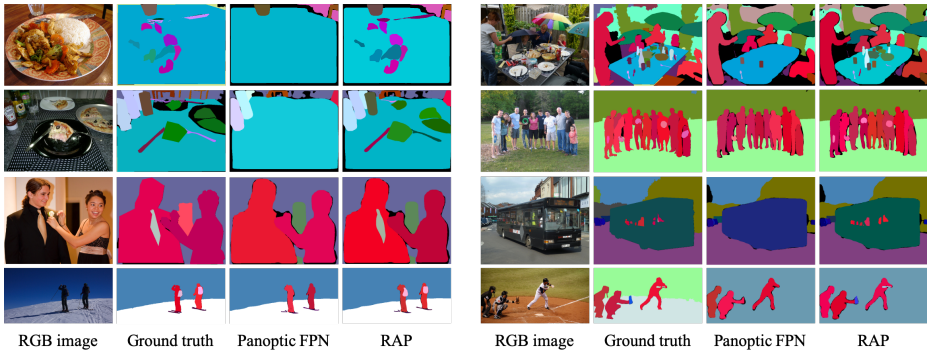

Fig. 6. Visualization on COCO val dataset.

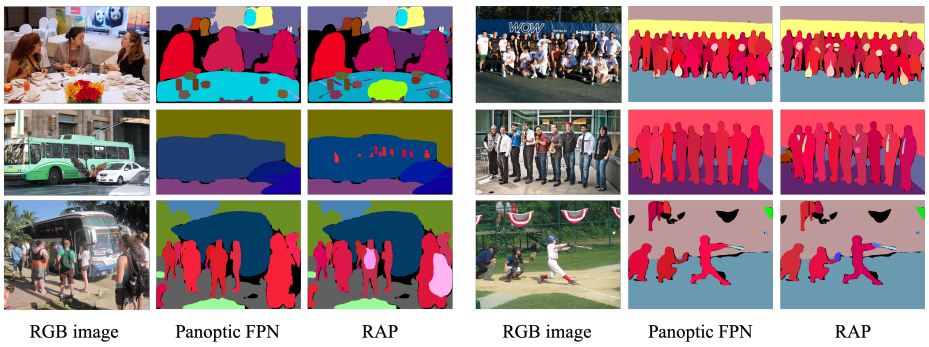

Fig. 7. Visualization on COCO test-dev dataset.

5 Conclusion

In this paper, we focus on tackling the overlap conflicts between instances during the fusion process in the existing panoptic segmentation methods. To this end, a relation aware panoptic segmentation network is proposed to learn a spatial relation value for every instance proposal. When overlaps occur, the relation value can be used instead of the confidence score to determine which instance is closer and assign the overlapping pixels to it. In addition, we analyze another conflict that can't be resolved only by using spatial relations, *i.e.*, excessive proposals conflict. Therefore, two overlapping area ratio check strategies are introduced to realize the non-maximum suppression by discarding useless proposals. We achieve significant improvement by generalizing RAP to a variety of backbones on COCO panoptic dataset, which verify the effectiveness of our model and fusion strategy.

In the future work, we plan to extend the spatial relation prediction from instances-wise to the pixel-wise to resolve the conflict between the instance and semantic branches.

References

1. Arnab, A., Torr, P.H.: Pixelwise instance segmentation with a dynamically instantiated network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 441–450 (2017)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
3. Bai, M., Urtasun, R.: Deep watershed transform for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5221–5229 (2017)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
6. Chen, W., et al.: Single-image depth perception in the wild. In: NeurIPS. pp. 730–738 (2016)
7. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3150–3158 (2016)
8. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
9. de Geus, D., Meletis, P., Dubbelman, G.: Panoptic segmentation with a joint semantic and instance segmentation network. arXiv preprint arXiv:1809.02110 (2018)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
12. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6409–6418 (2019)
13. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6399–6408 (2019)
14. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9404–9413 (2019)
15. Kirillov, A., Levinkov, E., Andres, B., Savchynskyy, B., Rother, C.: Instancecut: from edges to instances with multicut. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5008–5017 (2017)
16. Lazarow, J., Lee, K., Tu, Z.: Learning instance occlusion for panoptic segmentation. arXiv preprint arXiv:1906.05896 (2019)
17. Li, J., Raventos, A., Bhargava, A., Tagawa, T., Gaidon, A.: Learning to fuse things and stuff. arXiv preprint arXiv:1812.01192 (2018)

18. Li, Q., Arnab, A., Torr, P.H.: Weakly-and semi-supervised panoptic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 102–118 (2018)
19. Li, Y., Chen, X., Zhu, Z., Xie, L., Huang, G., Du, D., Wang, X.: Attention-guided unified network for panoptic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7026–7035 (2019)
20. Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.: Fully convolutional instance-aware semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2359–2367 (2017)
21. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
23. Liu, H., Peng, C., Yu, C., Wang, J., Liu, X., Yu, G., Jiang, W.: An end-to-end network for panoptic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6172–6181 (2019)
24. Liu, S., Jia, J., Fidler, S., Urtasun, R.: Sgn: Sequential grouping networks for instance segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3496–3504 (2017)
25. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8759–8768 (2018)
26. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
27. Massa, F., Girshick, R.: maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. `https://github.com/facebookresearch/maskrcnn-benchmark` (2018), accessed: [Insert date here]
28. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
29. Pinheiro, P.O., Collobert, R., Dollár, P.: Learning to segment object candidates. In: Advances in Neural Information Processing Systems. pp. 1990–1998 (2015)
30. Pinheiro, P.O., Lin, T.Y., Collobert, R., Dollár, P.: Learning to refine object segments. In: European Conference on Computer Vision. pp. 75–91. Springer (2016)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
32. Tighe, J., Niethammer, M., Lazebnik, S.: Scene parsing with object instances and occlusion ordering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3748–3755 (2014)
33. Tu, Z., Chen, X., Yuille, A.L., Zhu, S.C.: Image parsing: Unifying segmentation, detection, and recognition. International Journal of computer vision **63**(2), 113–140 (2005)
34. Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., Urtasun, R.: Upsnet: A unified panoptic segmentation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8818–8826 (2019)

35. Yang, T.J., Collins, M.D., Zhu, Y., Hwang, J.J., Liu, T., Zhang, X., Sze, V., Papandreou, G., Chen, L.C.: Deeperlab: Single-shot image parser. arXiv preprint arXiv:1902.05093 (2019)
36. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
37. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 405–420 (2018)
38. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)