REFINE: Prediction Fusion Network for Panoptic Segmentation

Jiawei Ren[*],[1] **Cunjun Yu**[*†],[1] **Zhongang Cai**[*],[1] **Mingyuan Zhang**,[1] **Chongsong Chen**[†],[1,2]
Haiyu Zhao[‡],[1] **Shuai Yi**,[1] **Hongsheng Li** [3]

[1]SenseTime Research
[2]Nanyang Technological University
[3]Multimedia Laboratory, The Chinese University of Hong Kong
{renjiawei, caizhongang, zhangmingyuan, zhaohaiyu, yishuai}@sensetime.com,
cunjun.yu@comp.nus.edu.sg, chen1129@e.ntu.edu.sg, hsli@ee.cuhk.edu.hk

Abstract

Panoptic segmentation aims at generating pixel-wise class and instance predictions for each pixel in the input image, which is a challenging task and far more complicated than naively fusing the semantic and instance segmentation results. Prediction fusion is therefore important to achieve accurate panoptic segmentation. In this paper, we present REFINE, pREdiction FusIon NEtwork for panoptic segmentation, to achieve high-quality panoptic segmentation by improving cross-task prediction fusion, and within-task prediction fusion. Our single-model ResNeXt-101 with DCN achieves PQ=51.5 on the COCO dataset, surpassing state-of-the-art performance by a convincing margin and is comparable with ensembled models. Our smaller model with a ResNet-50 backbone achieves PQ=44.9, which is comparable with state-of-the-art methods with larger backbones.

Introduction

As a step towards human-level visual scene understanding, the newly proposed task, panoptic segmentation (Kirillov et al. 2019b) has been drawing increasing attention in recent years. Different from instance segmentation, which focuses only on countable foreground instances, and semantic segmentation, which focuses on regions without differentiating instances, panoptic segmentation aims to provide a complete scene segmentation map that has both instance-wise labels for foreground objects and pixel-wise labels for background regions.

Under this holistic scene parsing setup, a *de-facto* multi-task strategy is generally adopted by state-of-the-art methods, in which both semantic segmentation and instance segmentation are conducted with two interacting networks. The semantic and instance segmentation results from both networks would be merged in the end to yield the desired panoptic segmentation.

Recently, many research works have exploited the benefits of the unified networks, with shared weights trained on both tasks, which show significant improvement over separately trained models (Li et al. 2019; Xiong et al. 2019; Yang et al. 2019) on either task. Moreover, to fully utilize the advantages of joint training, some works proposed to fuse instance information and semantic information by introducing designated fusion modules (Li et al. 2019, 2018). In light of this strategy, notable progress has been made in addition to the unified networks, where only the backbone is shared by both segmentation branches.

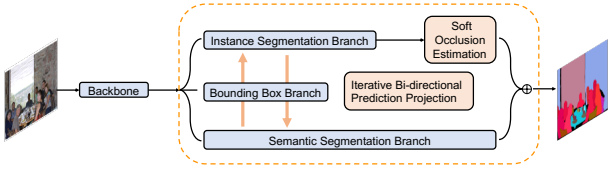

Figure 1: The proposed Prediction Fusion Network for Panoptic Segmentation

However, there is an inherent challenge for obtaining the final panoptic segmentation map: the fusion of predictions by semantic and instance segmentation. The challenge is of two-fold: 1) Cross-task synergy. Intrinsically, instance segmentation prediction head focuses on fine-grained instance details, while semantic segmentation prediction head is more attentive to holistic context. The fusion of the diverse predictions remains challenging in existing panoptic segmentation pipelines. Instance prediction and semantic prediction generally fail to reach a consensus. Existing works (Fu, Berg, and Berg 2019; Chen et al. 2018a; Li et al. 2018) only perform one-off uni-directional enhancement that uses one task branch to improve another, not fully exploiting the mutual benefits. 2) Inter-instance prediction fusion. To convert multiple instance segmentation masks to a single canvas in the end to generate the final segmentation map, it is necessary to have an accurate estimation of the spatial relation between pairs of instances. The fusion is especially crucial since an incorrect occlusion estimation would lead to missing objects or incomplete masks in the final segmentation map. The complex occlusion scenarios further exacerbate this issue. Existing works (Lazarow, Lee, and Tu 2020; Yang et al. 2020) leverage appearance and shape priors of predicted instance mask but neglect the mask quality associ-

[*]Equal contribution
[†]Work done at SenseTime Research
[‡]Corresponding author

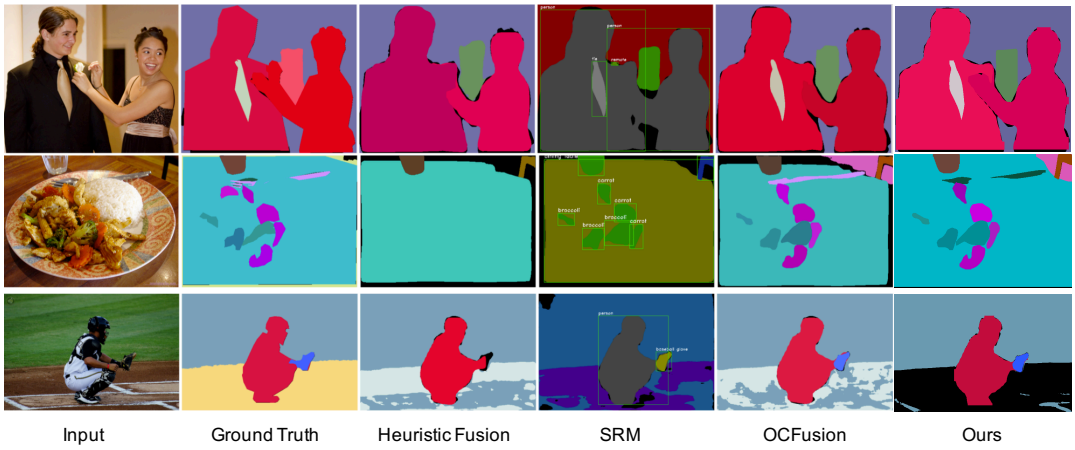

Figure 2: Qualitative results of the proposed and state-of-the-art panoptic segmentation methods. We compare with three methods, Heuristic Fusion (Kirillov et al. 2019a), Spatial Ranking Module (Liu et al. 2019) and OCFusion (Lazarow, Lee, and Tu 2020). Our proposed method shows significant improvements, there are better consistency between instance and semantic segmentation (row 1, less gap between semantic segmentation mask and instance segmentation mask, shown as black regions around people), less erroneous occlusion estimation (row 2, the spoon is successfully predicted to be occluded at the middle), and less false positive prediction (row 3, the less confident area is predicted have a continuous "unknown" label with much fewer false positive "holes")

ated with the bounding box confidence.

To tackle the aforementioned issues, we propose a novel panoptic segmentation pipeline, named REFINE, to harness the full potential of the unified framework. It optimizes the prediction fusion process to achieve high-quality panoptic segmentation. Our main contributions are:

- we present the Iterative Bi-directional Prediction Projection module, which not only refines the segmentation performance of each task by enriching the feature representations with the counterpart task's predictions, but also creates back-propagation routes from one task's supervision to the other task's prediction head, harmonizing the cross-task predictions.
- we propose the Inter-instance Soft Occlusion Estimation module to refine inter-instance occlusion prediction in different challenging scenarios, which exploits the bounding box confidences for more accurate and adaptive instance occlusion estimation.

Related Work

With the rapid development of deep-learning-based methods, research works on scene understanding, including object detection, semantic segmentation, and instance segmentation, have made remarkable progress. Despite the significant improvement in every single task, none of them is capable of providing pixel-level accurate labels to all foreground instances and background. Therefore, as a step into a higher-level scene understanding, panoptic segmentation was proposed in (Kirillov et al. 2019b) to bridge the gap.

Semantic Segmentation. As one of the most important scene understanding tasks, semantic segmentation serves as a fundamental component in computer vision. The very first method utilizing deep learning to tackle such a problem, fully convolutional network (FCN) (Shelhamer, Long, and Darrell 2014), unveils the power of the convolutional neural network in semantic segmentation. Since then, the encoder-decoder architecture is vastly adopted by the following works. Methods including UNet (Ronneberger, Fischer, and Brox 2015), DeepLab series (Chen et al. 2018b), DenseASPP (Yang et al. 2018) and PSPNet (Zhao et al. 2017) further pushed the boundary of semantic segmentation by leveraging contextual information. However, the task of semantic segmentation does not differentiate individual instances and thus leaves the scene segmentation incomplete.

Instance Segmentation. Unlike semantic segmentation, instance segmentation pays attention to the foreground instances rather than background regions. It aims to segment out each instance with both class labels and separate instance labels. There exist two main categories of methods for instance segmentation. One intuitive way is to predict pixel-wise label based on semantic segmentation (Arnab and Torr 2017; Liang et al. 2017; Liu et al. 2018b). More recent works, such as Mask-RCNN (He et al. 2017), PANet (Liu et al. 2018a), take advantage of the bounding boxes to generate accurate instance masks. By jointly training both segmentation and detection branches, the unified models yield better results for both tasks. Due to the nature of proposal-based instance segmentation, background segmentation is not involved in the prediction. In HTC (Chen et al. 2019), experiment results indicate that both semantic and instance segmentation benefit from the joint training of both tasks, making it possible to extend the method into panoptic segmentation.

Panoptic Segmentation. Panoptic segmentation (Kirillov et al. 2019b) unifies the tasks of semantic segmentation and instance segmentation to assign each pixel a cate-

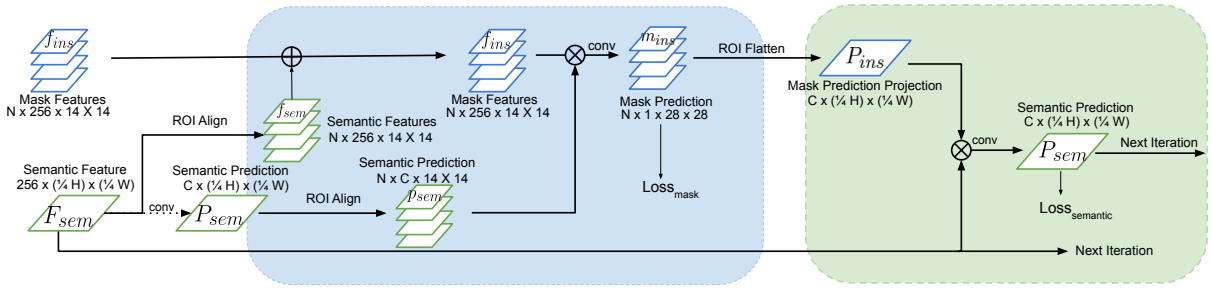

Figure 3: The proposed Iterative Bi-directional Prediction Projection (IBPP). $\oplus$ denotes element-wise sum and $\otimes$ denotes concatenation of features or predictions. Initial semantic predictions are first generated before the bi-directional projection starts, which can repeat multiple times. The blue-shaded area indicates the projection of semantic segmentation predictions onto instance segmentation; the green-shaded area indicates the projection of instance segmentation prediction onto semantic segmentation. IBPP fosters synergy between two segmentation task branches in a recursive manner

gory label and to segment each foreground object instance simultaneously. Panoptic FPN (Kirillov et al. 2019a) uses a single backbone network, followed by Feature Pyramid Network (FPN (Lin et al. 2017)) and multiple heads in both instance and semantic branches. (Li et al. 2019) proposed AUNet, which integrates attention modules guided by the instance branch to improve the background segmentation results. UPSNet (Xiong et al. 2019) adopted a unified architecture to predict both semantic and instance confidence scores for each pixel.

Instance Occlusion Estimation. Instance occlusion estimation is important in panoptic segmentation. In early methods, the occlusion relation was inferred from each instance's classification score. This heuristic does not provide accurate occlusion estimation. OANet (Liu et al. 2019) predicted spatial ranking scores and solved the mask overlaps, whose results are in favor of instances with higher scores. OCFusion (Lazarow, Lee, and Tu 2020) proposed an appearance-based occlusion head to learn the occlusion relation between pairs of instances from their cropped feature maps and masked bitmap. Similarly, SOGNet (Yang et al. 2020) proposed a module to utilize information including bounding boxes, class labels, and masked bitmap to build an occlusion relation graph.

Proposed Method

To better fuse the predictions from both semantic segmentation and instance segmentation tasks, our unified panoptic segmentation framework train instance segmentation and the semantic segmentation in a joint manner (see in Figure 1). We propose novel bi-directional prediction projection and instance-prediction occlusion estimation modules, which effectively tackles the semantic-instance prediction fusion and inter-instance prediction fusion, respectively.

Iterative Bi-directional Prediction Projection

Although existing panoptic segmentation frameworks have well exploited the advantages of feature sharing, the instance segmentation predictions and semantic segmentation predictions are still separately obtained, which inevitably requires further processing to be effectively fused. We propose the Iterative Bi-directional Prediction Projection (Iterative BPP) to utilize the predictions of both semantic and instance segmentation to iteratively assist the feature learning of each other (see in Figure 3). We validate that our method is compatible with common operations (such as ROI Align (He et al. 2017) and ROI Flatten (Fu, Berg, and Berg 2019)) and it does not require specifically designed components that are tailored to our purpose.

In the general pipeline of panoptic segmentation, there are an instance segmentation head H_{ins} and a semantic segmentation head H_{sem} following the FPN structure: H_{ins} takes in instance feature $f_{\text{ins}} \in \mathbb{R}^{256\times14\times14}$ as inputs and outputs instance masks $m_{\text{ins}} \in \mathbb{R}^{1\times28\times28}$; H_{sem} takes in semantic features $F_{\text{sem}} \in \mathbb{R}^{256\times(H/4)\times(W/4)}$ and outputs semantic predictions $P_{\text{sem}} \in \mathbb{R}^{C_{\text{sem}}\times(H/4)\times(W/4)}$, i.e.,

$$m_{\text{ins}} = H_{\text{ins}}(f_{\text{ins}}), \quad P_{\text{sem}} = H_{\text{sem}}(F_{\text{sem}}). \tag{1}$$

In our framework, we further extend the semantic head H_{sem} to predict both C_{ins} foreground classes and C_{sem} background classes, so that the predictions include both classes for instance segmentation and semantic segmentation, i.e., $P_{\text{sem}} \in \mathbb{R}^{(C_{\text{sem}}+C_{\text{ins}})\times(H/4)\times(W/4)}$.

In our proposed Iterative BPP, an initial semantic prediction is first generated. After that, we iteratively project semantic features and predictions to instance features to assist instance segmentation performance and then project instance predictions to semantic features for assisting semantic segmentation performance. The iterations repeat and can propagate useful information between the two tasks, and gradually encourage them to reach consensus on the final segmentation results. Our Iterative BPP is illustrated in Fig. 3, and has the following steps in each of the iterations:

1. In the very first iteration, we adopt the conventional semantic segmentation head H_{sem} to obtain the initial semantic prediction $P^{(0)}_{\text{sem}}$.
2. At each iteration t, given foreground instance boxes, we ROI Align semantic features F_{sem} and semantic predictions P_{sem} at the RoIs of the previous iteration $t-1$ as

$$f_{\text{sem}} = \text{ROI-Align}(F_{\text{sem}}), \tag{2}$$

$$p^{(t-1)}_{\text{sem}} = \text{ROI-Align}\left(P^{(t-1)}_{\text{sem}}\right) \tag{3}$$

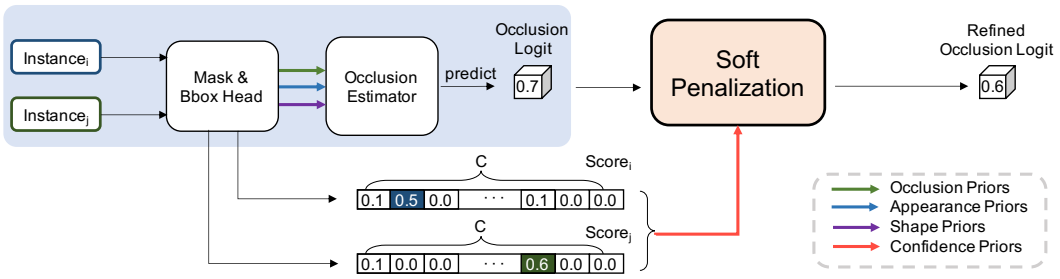

Figure 4: Our proposed soft occlusion estimation module. The blue-shaded area is the conventional occlusion estimation pipeline. In contrast, we utilize an additional prior, confidence prior, to further enhance accurate occlusion relation estimation

The ROI Aligned semantic features f_{sem} capture semantic information of the semantic segmentation but at the instance locations. We propagate the semantic features f_{sem} to the instance branch and sum them together to obtain the enhanced instance features $f_{\text{ins}} + f_{\text{sem}}$.

In addition, the ROI Aligned semantic predictions $p_{\text{sem}}^{(t-1)}$ are also propagated to the instance branch and concatenated with the enhanced instance features to refine the instance prediction as

$$m_{\text{ins}}^{(t)} = \tilde{H}_{\text{ins}}\left(\left[(f_{\text{ins}} + f_{\text{sem}}); p_{\text{sem}}^{(t-1)}\right]\right), \tag{4}$$

where $[\cdot;\cdot]$ denotes concatenation and $\tilde{H}_{\text{ins}}$ is an instance segmentation head. In this way, the segmentation branch propagates features and predictions to assist the learning of instance segmentation.

3. The refined instance predictions $m_{\text{ins}}^{(t)}$ are then propagated to the semantic branch for refining the semantic segmentation. We first adopt ROI Flatten (Li et al. 2018) to pool all instance masks of size $1 \times 28 \times 28$ onto a single full-channel canvas $P_{\text{ins}} \in \mathbb{R}^{C_{\text{ins}} \times (H/4) \times (W/4)}$ as

$$P_{\text{ins}}^{(t)} = \text{ROI-Flatten}\,(m_{\text{ins}}^{(t)}). \tag{5}$$

The newly projected instance-based canvas $P_{\text{ins}}^{(t)}$ is then concatenated with the semantic feature map F_{sem} and input into the new semantic segmentation head $\tilde{H}_{\text{sem}}$

$$P_{\text{sem}}^{(t)} = \tilde{H}_{\text{sem}}\left(\left[P_{\text{ins}}^{(t)}; F_{\text{sem}}\right]\right). \tag{6}$$

4. The above steps 2 and 3 can iteratively refine the instance segmentation and semantic segmentation predictions, $m_{\text{ins}}^{(t)}$ and $P_{\text{sem}}^{(t)}$, in an alternative manner.

Comparing with methods only use the cross-task feature to repeatedly refine the predictions (Chen et al. 2019), our Iterative BPP has mainly two merits: (1) We for the first time propose to propagate cross-task predictions as a stronger prior, in addition to simple cross-task features, for prediction refinement of the two tasks; (2) the cross-task propagated predictions would receive more routes of gradients in the loops and thus help to learn more discriminative features.

Different from existing uni-directional method in instance segmentation (Fu, Berg, and Berg 2019) and semantic segmentation (Chen et al. 2018a), our bi-directional projection enables iterative refinement that generates more routes for back-propagation gradients to flow, and thus enables better feature learning. Other than forcing the predictions of the two tasks to be consistent by adding an L2 regularization term (Li et al. 2018), the Iterative BPP achieves unifying cross-task predictions in an implicit and easy-to-optimize manner.

Instance Prediction with Soft Occlusion Estimation

Given the semantic prediction and instance prediction, we need to combine mask predictions of the two types to generate the panoptic segmentation map, which is generally achieved by pasting instance masks onto the semantic segmentation map one at a time according to the inter-instance occlusion relations. Such occlusion relations are therefore of great importance to obtain the high-quality panoptic segmentation map.

Existing works (Lazarow, Lee, and Tu 2020; Yang et al. 2020) consider the following three types of important prior information when estimating occlusion relation between pairs of instance predictions:

- Inter-class occlusion priors. Given instances of two different classes, instances of one class might be more likely to occlude instances of the other class. For instance, for a tie instance and a person instance in the same image, the tie is more likely to occlude the person. This is because if the tie, a relatively small instance, is occluded by the person, it might not be captured by the camera at all.
- Inter-instance appearance priors. Given a pair of occluding/occluded instances, their appearances can provide strong cues on estimating their occlusion relation.
- Inter-instance shape priors. Given a pair of occluding/occluded instances, the instance in front should obtain a more complete mask compared with the one being occluded, which might only have a partial mask. For instance, a complete car prediction mask is more likely to occlude a person with partial shape.

However, we argue that the abovementioned priors neglect the quality of the bounding boxes (confidence scores). Hence, we propose an additional prior:

- Inter-instance confidence priors. Given a pair of instances, the instance in the front generally has a more complete ap-

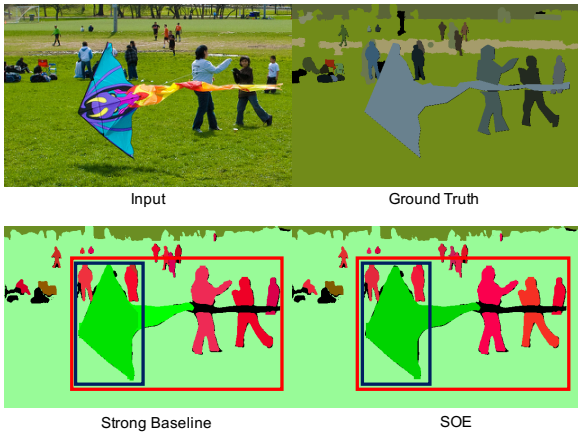

Figure 5: Confidence prior is crucial for occlusion estimation. Multiple instance masks can be predicted for the same instance. More complete instance masks are usually associated with higher confidence scores. For instance, for a kite-kite pair, the kite with higher confidence is more likely to occlude the one with lower confidence. Compared to the strong baseline, SOE leverages all priors (including the confidence prior) and predicts the better mask to be placed in the front

pearance than the one in the back. The instance in front is therefore more likely to have higher classification scores. We explain in Figure 5 that confidence prior is essential for high quality occlusion estimation.

We design a soft penalization that leverages all four priors. We therefore propose the Soft Occlusion Estimation (SOE, see in Figure 4) to make an accurate estimation of pairwise inter-instance occlusion estimation. The estimated occlusion confidences can be later used to determine the instance predictions' pasting order and to also modulate their class scores.

A Strong Baseline. We first follow (Lazarow, Lee, and Tu 2020; Yang et al. 2020) and model the occlusion estimation as a neural network, which takes appearance features, predicted class labels of a pair of overlapping instances i and j, instance mask predictions as inputs and outputs an occlusion confidence score between $[0, 1]$. More implementation details are included in the supplementary material.

Soft Penalization. Based on the estimated inter-instance occlusion score $O(i, j)$ between a pair of instances i and j, we further modulate the occlusion confidence $O(i, j)$ and multiply it with the original maximal class score of each instance

$$\tilde{O}(i,j) = s_i \exp\left(-\frac{(O(i,j)-1)^2}{\sigma}\right), \tag{7}$$

where s_i and s_j are the maximal class scores out of C classes for instances i and j, respectively. The above soft-weighting scheme term takes inter-instance confidence priors into account. As instances with higher class confidences s_i are more likely to have complete appearances and masks, and are more likely to occlude other masks. Hence, it should be assigned a higher occlusion confidence. The Gaussian penalty term $\exp(-(O(i,j)-1)^2/\sigma)$ increases gradually when the occlusion confidence is low, while it shows more significant effect when the occlusion confidence is closer to one. This property is proven to be beneficial in (Bodla et al. 2017). The hyperparameter σ adjusts the impact of original occlusion prediction logit $O(i, j)$'s impact to the final confidence. We iterate over all overlapping pairs to estimate their occlusion relations. If $\tilde{O}(i,j) > \tilde{O}(j,i)$, the instance i is considered to occlude instance j and should be pasted over instance j's mask in the final segmentation map.

Semantic Prediction Refinement with Unknown Erasing

It is better to indicate uncertainty than to give wrong prediction. This is reflected by Panoptic Quality, the accuracy metric of panoptic segmentation

$$\text{PQ} = \underbrace{\frac{\sum_{(p,g)\in TP}\text{IoU}(p,g)}{|TP|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}}. \tag{8}$$

which penalizes an unknown class prediction (0.5FN) less than an incorrect class prediction (0.5FN + 0.5FP). We introduce Unknown Erasing (UE) to further refine the semantic predictions before generating the panoptic segmentation map. The UE first computes the pixel-wise average classification confidences for each continuous region (which we name as connected components (*CC*)) in the semantic prediction map. Areas whose average scores are below a threshold are erased. Since small fragments with abnormally high-confidence scores would not be erased and still contributes to false positives, we dilate each *CC* to eliminate the isolated small regions and compute the dilated regions' scores for erasing. The pseudo-code for UE can be found in the supplementary material.

Experiments

In this section, we evaluate our approach on COCO dataset (Lin et al. 2014) for panoptic segmentation. More experiments (including those on Cityscapes (Cordts et al. 2016)) can be found in the supplementary material due to space constraint. Comparisons with state-of-the-art methods demonstrate the effectiveness of our overall framework. We also evaluate each sub-module of our method together with a detailed analysis. We further decompose our method to find out the contribution of each component. Qualitative results are shown in Figure 2.

Dataset and Evaluation metric. Evaluation is performed on COCO2017 dataset (Lin et al. 2014). We take the default 118k/5k/20k split for train/val/test from COCO2017. For experiments related to a single component, we report panoptic segmentation results on the validation set. For the overall results, we report the panoptic segmentation performance on the test set from the official split.

Implementation Details. We choose the off-the-shelf Faster-RCNN with FPN and ResNet-50 backbone with Group Normalization (Wu and He 2018) as our object detector. Details can be found in the supplementary material.

Method	Backbone	Image Size	PQ	PQ^{th}	PQ^{st}	SQ	RQ	Split
UPSNet† (Xiong et al. 2019)	ResNet-50	800 × 1333	42.5	48.5	33.4	78.0	52.4	*val*
Panoptic-FPN (Kirillov et al. 2019a)	ResNet-101	800 × 1300	40.9	48.3	29.7	-	-	*test-dev*
OANet (Liu et al. 2019)	ResNet-101	800 × 1333	41.3	50.4	27.7	-	-	*test-dev*
AUNet (Li et al. 2019)	ResNeXt-152-DCN	MS + Flip	46.5	55.9	32.5	81.0	56.1	*test-dev*
UPSNet (Xiong et al. 2019)	ResNet-101-DCN	MS + Flip	46.6	53.2	36.7	80.5	56.9	*test-dev*
OCFusion (Lazarow, Lee, and Tu 2020)	ResNeXt-101-DCN	MS + Flip	46.7	54.0	35.7	-	-	*test-dev*
SOGNet (Yang et al. 2020)	ResNet-101-DCN	MS + Flip	47.8	-	-	80.7	57.6	*test-dev*
UTIPS (Li, Qi, and Torr 2020)	ResNet-101-DCN	800 × 1333	47.2	53.5	37.7	81.1	57.2	*test-dev*
DetectoRS (Qiao, Chen, and Yuille 2020)	ResNeXt-101-DCN*	MS + Flip	49.6	57.8	37.1	-	-	*test-dev*
Ours	ResNet-50	800 × 1333	44.9	51.7	34.7	79.4	54.2	*val*
Ours	ResNet-101-DCN	800 × 1333	49.6	57.5	37.7	81.7	59.3	*test-dev*
Ours	ResNeXt-101-DCN	800 × 1333	**51.5**	**59.6**	**39.2**	**82.6**	**61.3**	*test-dev*

Table 1: Comparison with state-of-the-art methods on COCO dataset with various backbones. MS + Flip refers to multi-scale test with horizontal flipping. † denotes deformable convolution adopted in semantic branch. * denotes that it uses stuff prediction from DeepLabv3+ with Wide-ResNet-41 as backbone. Our method outperforms the state-of-the-art methods with convincing margins. Note that our method does not require any test-time augmentation. Please see the supplementary material for Cityscapes (Cordts et al. 2016) results

Method	Backbone	PQ	PQ^{th}	PQ^{st}	AP^{b}	AP^{m}	FPS
Panoptic-FPN (Kirillov et al. 2019a)	ResNet-101	40.9	48.3	30.0	-	-	10.0
Cascade Panoptic-FPN	ResNet-101-DCN	46.1	-	-	47.4	41.3	7.3
Ours	ResNet-101-DCN	49.3	57.0	37.7	48.0	41.8	5.5

Table 2: Additional Comparisons on COCO *val*. Cascade Panoptic-FPN includes deformable convolutions and Cascade-RCNN (Cai and Vasconcelos 2018), on top of the original Panoptic-FPN for better bounding box localization and mask prediction. Interestingly, our method shows a significant improvement on PQ with a marginal improvement on AP. This shows that the semantic segmentation and effective prediction fusion, instead of better bounding box prediction, contribute towards a better panoptic segmentation. Moreover, although our method is designed for high performance, it still achieves a decent inference speed (0.7x Cascade Panoptic-FPN)

Comparison with the State of the Arts

In this section, we compare our model with state-of-the-art methods on panoptic segmentation. As shown in Table 1, we report our result with three different setups to make a fair comparison on COCO. With a plain ResNet-50 as the backbone, our approach has exceeded most of the existing methods. To validate our method works with a heavier network, we then employ ResNeXt (Xie et al. 2017), as the backbone with deformable convolution (Dai et al. 2017). The model reaches PQ of 51.5, which exceeds all the state-of-the-art single models by clear margins. No test-time augmentation is adopted in our proposed method for all the results reported in the table.

We further compare our method with a stronger target, Cascade Panoptic-FPN[1], which adds deformable convolutions and Cascade-RCNN (Cai and Vasconcelos 2018) to the original Panoptic-FPN in order to obtain better bounding box localization and mask prediction. Cascade Panoptic-FPN is trained for an even longer schedule, 270k iterations while we train our model for 180k iterations under the same setup. Interestingly, Table 2 shows our model achieves much higher Panoptic Quality than Cascade Panoptic-FPN with a limited advantage in bounding box AP. This highlights that better semantic mask prediction and prediction fusion contribute more to panoptic segmentation performance than a better bounding box localization. This observation is in line with (Xiong et al. 2019). Moreover, we find that our method can still achieve a decent inference speed.

Component Analysis

In this section, we use various experiments to analyze the contribution of each component proposed in this paper, as shown in Table 3. Our baseline is built upon a Panoptic-FPN (Kirillov et al. 2019a) with 39.6 PQ. In Table 3, we show that each of our proposed modules contributes to an overall improvement: IBPP improves both instance and semantic segmentation, SOE addresses occlusion issue for instance segmentation and UE refines semantic segmentation results.

We take Panoptic-FPN with ResNet-50 as our baseline. Since the effect of occlusion signifies with higher instance segmentation quality, we then use the output of a model trained with ResNeXt-101 as the input of Soft Occlusion Estimation.

Iterative Bi-directional Prediction Projection (IBPP). IBPP is dedicated to propagating the predictions across tasks to improve both ends. We perform three iterations of IBPP, the same as Cascade-RCNN (Cai and Vasconcelos 2018), which has been proved efficient in object detection.

In Table 4, we take a plain Panoptic-FPN (Kirillov et al. 2019a) model with ResNet-50 and ResNet-101 as our baseline. By introducing bi-directional prediction projection, predictions from the semantic branch and instance branch

[1] implemented in *detectron2* with its version at commit 8999946

IBPP	SOE	UE	PQ	PQ^{th}	PQ^{st}
-	-	-	39.6	46.1	29.8
✓	-	-	42.2 (+2.6)	48.4	32.9
✓	✓	-	44.2 (+2.0)	51.7	32.9
✓	✓	✓	44.9 (+0.7)	51.7	34.7

Table 3: Component Analysis. IBPP: Iterative Bi-directional Prediction Projection. SOE: Soft Occlusion Estimation. UE: Unknown Erasing. IBPP is designed to enhance cross-task interaction that benefits both instance and semantic segmentation; SOE tackles occlusion and improves instance segmentation; UE removes low confidence semantic mask prediction. The experiment results observed justifies the effectiveness of our design

Method	Iteration	Backbone	PQ	PQ^{th}	PQ^{st}
Baseline	1	ResNet-50	39.6	46.1	29.8
IBPP	1	ResNet-50	41.0	47.5	30.7
IBPP	2	ResNet-50	41.6	48.0	31.5
IBPP	3	ResNet-50	42.2	48.4	32.9
Baseline	1	ResNet-101	40.9	48.3	30.0
IBPP	3	ResNet-101	43.4	49.4	34.4

Table 4: Results of a baseline (Panoptic-FPN (Kirillov et al. 2019a)) and IBPP. AP for both bounding box and instance segmentation are shown in COCO style. A single iteration of IBPP is already effective, showing thatsegmentation task branches benefit from each other; more iteration further improves the performance, demonstrating the effectiveness of cascaded refinement. We cap the number of iterations at three to match the setting of Cascade-RCNN(Cai and Vasconcelos 2018) but more iterations are possible. Moreover, IBPP gives significant improvements regardless of the backbone used

end up with better consistency (Figure 2). With different backbones, IBPP leads corresponding baseline by 2.6 PQ for ResNet-50 and 2.5 PQ for ResNet-101. Both PQ^{th} and PQ^{st} are benefited significantly from such prediction-level fusion.

We also ablate the effect of the different number of iterations in IBPP in Table 4. IBPP outperforms the baseline by 1.4 PQ when only one iteration of refinement is applied. With iterative refinements, the performance of IBPP progressively improves. Notice that in this section, Soft Occlusion Estimation is not involved and only IBPP is added to the baseline model.

Soft Occlusion Estimation (SOE). In this section, we demonstrate that a better inter-instance prediction fusion can be achieved with the proposed SOE. An IBPP model with a ResNeXt-101 backbone serves as our baseline in this section. We reproduced four popular occlusion estimation methods on our baseline: heuristic fusion (Kirillov et al. 2019b), SHR (Wang et al. 2019), SRM (Liu et al. 2019), and OCFusion (Lazarow, Lee, and Tu 2020) . As shown in Table 5, our proposed occlusion estimator outperforms SHR, SRM and OCFusion. Compared with OCFusion which only utilizes mask prediction and mask feature as inputs, we show

Method	σ	PQ^{th}	SQ^{th}	RQ^{th}
Heuristic Fusion	-	54.8	83.9	57.6
SHR	-	56.9	84.1	67.2
SRM	-	57.1	83.6	67.8
OCFusion	-	58.4	83.9	69.2
Strong Baseline	-	58.5	83.9	69.3
SOE	2.0	58.8	84.1	69.5
SOE	**2.5**	**58.9**	**84.2**	**69.5**
SOE	4.0	58.6	84.2	69.1

Table 5: Comparison of our proposed Soft Occlusion Estimation module with existing occlusion estimation solutions and analysis on the hyperparameter σ. We empirically show that $\sigma = 2.5$ gives the optimal result. Our method outperforms SHR (Wang et al. 2019), SRM (Liu et al. 2019) and OCFusion (Lazarow, Lee, and Tu 2020)

Mask Pred	Mask Feat	Label	σ	PQ^{th}	SQ^{th}	RQ^{th}
-	-	-	-	54.8	83.9	57.6
✓	-	-	-	57.4	83.8	67.8
✓	✓	-	-	58.4	83.9	69.2
✓	✓	✓	-	58.5	83.9	69.3
✓	✓	✓	✓	58.9	84.2	69.5

Table 6: Ablation study of confidence prior. Mask Prediction: shape prior; Mask Feature: appearance prior; Label: class prior; $\boldsymbol{\sigma}$: confidence prior; The first row shows a heuristic baseline. We follow (Lazarow, Lee, and Tu 2020; Yang et al. 2020) and build a strong baseline (row 4) that utilizes shape, appearance and class priors. We then show that our soft penalization that utilizes confidence prior can further improve the performance

that our approach makes use of all available informative cues and result in more accurate occlusion estimations. A soft-weighting scheme is introduced to fully utilize the power of different cues. We weight the occlusion prediction with different hyper-parameter σ and the optimum is archived with $\sigma = 2.5$. We show in Table 5 that the performance improved by 0.4 PQ^{th} after we adopted the soft-weighting scheme. We show a component analysis of SOE in Table 6, which indicates that confidence prior is complementary to other priors (occlusion, appearance, and shape priors), and leads to further improvement of panoptic segmentation performance.

Conclusion

We present an effective prediction fusion network to achieve high-accuracy panoptic segmentation. The proposed method consists of two modules, Iterative Bi-directional Prediction Projection for better interaction between the two task branches, and Soft Occlusion Estimation for robust spatial relation prediction. Our unified framework achieves state-of-the-art performance on the COCO dataset. As the future work, we aim to make our method faster and lighter.

Acknowledgments

This work is supported in part by the General Research Fund through the Research Grants Council of Hong Kong under

Grants (Nos. CUHK14208417, CUHK14207319), in part by the Hong Kong Innovation and Technology Support Program (No. ITS/312/18FX), in part by CUHK Strategic Fund.

References

Arnab, A.; and Torr, P. H. 2017. Pixelwise instance segmentation with a dynamically instantiated network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 441–450.

Bodla, N.; Singh, B.; Chellappa, R.; and Davis, L. S. 2017. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, 5561–5569.

Cai, Z.; and Vasconcelos, N. 2018. Cascade R-CNN: Delving into High Quality Object Detection. In *CVPR*.

Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. 2019. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4974–4983.

Chen, L.-C.; Hermans, A.; Papandreou, G.; Schroff, F.; Wang, P.; and Adam, H. 2018a. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4013–4022.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018b. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *ECCV*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Fu, C.-Y.; Berg, T. L.; and Berg, A. C. 2019. Imp: Instance mask projection for high accuracy semantic segmentation of things. In *Proceedings of the IEEE International Conference on Computer Vision*, 5178–5187.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Kirillov, A.; Girshick, R.; He, K.; and Dollár, P. 2019a. Panoptic feature pyramid networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6399–6408.

Kirillov, A.; He, K.; Girshick, R.; Rother, C.; and Dollár, P. 2019b. Panoptic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9404–9413.

Lazarow, J.; Lee, K.; and Tu, Z. 2020. Learning Instance Occlusion for Panoptic Segmentation. *CVPR* abs/1906.05896.

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

Li, Q.; Qi, X.; and Torr, P. H. 2020. Unifying training and inference for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13320–13328.

Li, Y.; Chen, X.; Zhu, Z.; Xie, L.; Huang, G.; Du, D.; and Wang, X. 2019. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7026–7035.

Liang, X.; Lin, L.; Wei, Y.; Shen, X.; Yang, J.; and Yan, S. 2017. Proposal-free network for instance-level object segmentation. *IEEE transactions on pattern analysis and machine intelligence* 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6172–6181.

Liu, S.; Qi, L.; Qin, H.; Shi, J.; and Jia, J. 2018a. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8759–8768.

Liu, Y.; Yang, S.; Li, B.; Zhou, W.; Xu, J.; Li, H.; and Lu, Y. 2018b. Affinity derivation and graph merge for instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 686–703.

Qiao, S.; Chen, L.-C.; and Yuille, A. 2020. DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. *arXiv preprint arXiv:2006.02334* .

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Shelhamer, E.; Long, J.; and Darrell, T. 2014. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* .

Wang, S.; Liu, T.; Liu, H.; Ma, Y.; Li, Z.; Wang, Z.; Zhou, X.; Yu, G.; Zhou, E.; Zhang, X.; et al. 2019. Joint COCO and Mapillary Workshop at ICCV 2019: Panoptic Segmentation Challenge Track Technical Report: Explore Context Relation for Panoptic Segmentation. In *ICCV workshop*.

Wu, Y.; and He, K. 2018. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Xiong, Y.; Liao, R.; Zhao, H.; Hu, R.; Bai, M.; Yumer, E.; and Urtasun, R. 2019. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8818–8826.

Yang, M.; Yu, K.; Zhang, C.; Li, Z.; and Yang, K. 2018. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3684–3692.

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* .

Yang, Y.; Li, H.; Li, X.; Zhao, Q.; Wu, J.; and Lin, Z. 2020. Sognet: Scene overlap graph network for panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12637–12644.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.