2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10222709

OVERLAP LOSS: RETHINKING WEAKLY SUPERVISED INSTANCE SEGMENTATION IN CROWDED SCENES

Shanghang Jiang[1], Shichao Zhao[2], Meng Wu[1,], Le Zhang[1], Feng Zhou[2]*

[1]School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China
[2]Algorithm Research, Aibee Inc.

ABSTRACT

Weakly supervised instance segmentation (WSIS) has gained increasing popularity in recent years due to low labelling cost. However, its performance deteriorates dramatically in more challenging crowded scenario, which is caused by overlapping among similar objects. To ameliorate the negative effects of instance overlapping, we propose a new loss, i.e., OverlapLoss, which achieves instance disentanglement between masks according to the degree of overlapping among instances. Besides, a new dataset of CrowdHuman Instance Segmentation (CIS) is presented to bridge the gap in crowded scenes. Experiments on the CIS and COCO datasets validate that the proposed loss can improve the baseline in typical crowed scenes by at least 2% and in uncrowded scenes by more than 0.3% w.r.t. absolute AP. The code and dataset are available at: https://github.com/shanghangjiang/CIS.

***Index Terms*—** Instance segmentation, weakly supervised instance segmentation, box supervision

1. INTRODUCTION

Instance segmentation aims to predict the pixel-wise masks and categories of instances of interest, which is one of the most fundamental tasks in computer vision. Although instance segmentation is able to provide more accurate and finer mask level object location than detection, it used to be blamed for burdensome pixel-wise mask annotations.

To deal with this problem, weakly supervised instance segmentation (WSIS) takes advantages of image-level [1, 2], point-level [3] or box-level annotations [4, 5, 6, 7, 8] rather than the pixel-wise mask labels. Among of these, box-supervised instance segmentation attempts to utilize bounding box of an instance and achieves superior performance, which has recently received more attention in view of its low labelling cost and concurrent development of object detection. Following the vein of its fully supervised counterpart, i.e., CondInst [9], BoxInst [6] is a milestone method which significantly outperforms previous weakly-supervised approaches by simply revising loss function to implement end-to-end instance segmentation. Other WSIS methods [4, 5, 7] can compare with and even outperform some fully supervised approaches [10, 11] in uncrowded scenes.

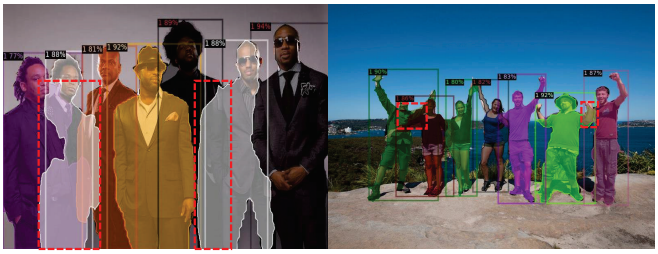

Fig. 1. Semantic ambiguity between neighboring instances in crowded scenes. Overlapping in inference by weakly supervised instance segmentation is largely due to the independent mask generation, as is shown in red dashed boxes.

However, WSIS methods still struggle with complex instance interaction in crowded scenes. Taking a common scenario of group photo for an example, even the state-of-the-art (SOTA) method fails to predict the overlapping area in crowded scenes, as is illustrated in Fig. 1. Overlapping among instances is thus supposed to be a critical issue for WSIS methods, since it incurs semantic ambiguity. Speaking of crowded scenarios, more works focus on object detection [12, 13, 14, 15] with a few on instance segmentation [16].

Unlike previous instance segmentation methods which generate masks in an independent way, we are the first to explore the overlapping constraint in crowed scenarios. Specifically, our contributions are two-fold as follows:

- From the perspective of instance overlapping, we introduce a novel OverlapLoss to enhance discrimination between overlapped instances. Experiments verify its superiority in both crowded and uncrowded scenes.
- To fill in the blanks of evaluation in crowded scene, we will release a new instance segmentation dataset by augmenting CrowdHuman [17].

*Corresponding author: Meng Wu, mail: wumeng@nwpu.edu.cn

This work was supported by the Natural Science Basic Research Plan in Shaanxi Province of China (2020JQ-208), and Key Research and Development Program of Shaanxi (2022GY-285, 2020SF-391).

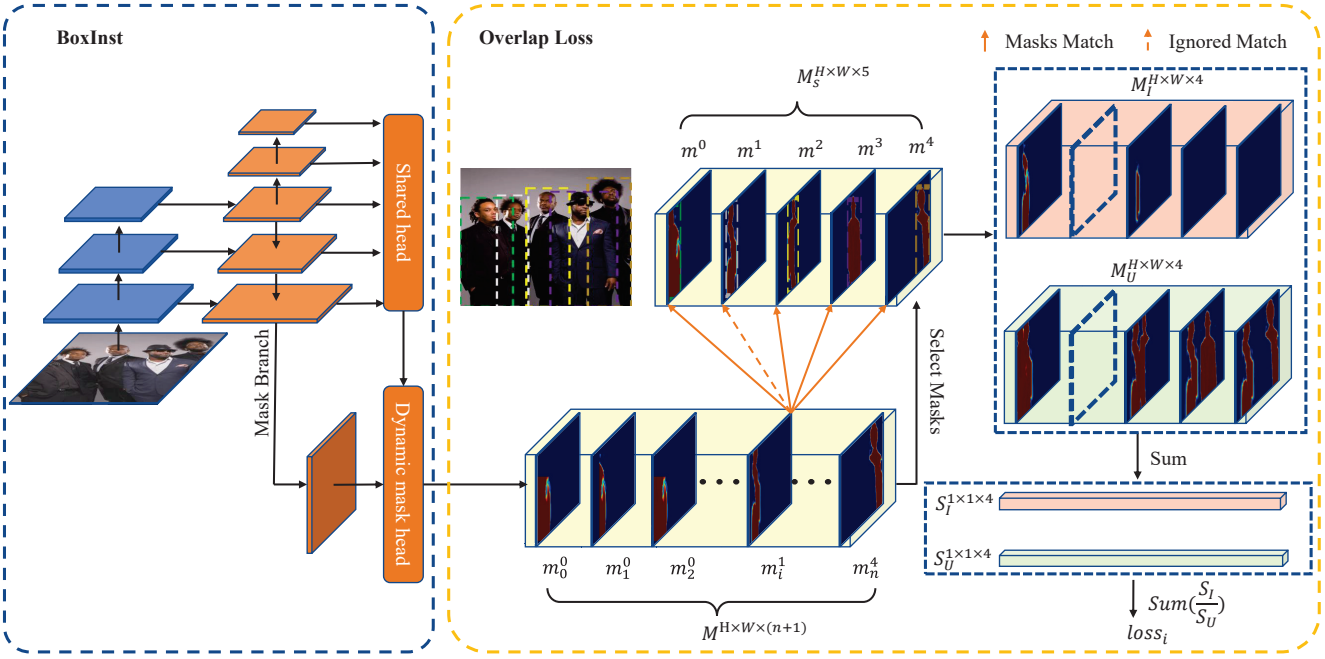

Fig. 2. The architecture of our approach. Blue dashed box indicates the backbone of any off-the-shelf box-supervised instance segmentation, e.g., BoxInst. Orange dashed box depicts the proposed OverlapLoss. S_I and S_U are the vector of intersection score and union score.

2. PROPOSED METHOD AND DATASET

2.1. Overall Architecture

As a SOTA method in box-supervised instance segmentation, BoxInst [6] is the framework we impose our loss on, which replaces the original pixel-wise mask losses in CondInst [9] with projection loss and affinity loss. The projection loss forces horizontal/vertical lines inside bounding boxes to predict at least one foreground pixel, while the affinity loss forces pixels with similar colors to have the same label. Despite its weak supervision, BoxInst [9] achieves an impressive 85% of fully-supervised performance on COCO [18]. Following [6], we append our OverlapLoss to the mask branch to deal with the semantic ambiguity, as Fig. 2 illustrates.

2.2. The Overlap Loss

In general, the mask generation of box-supervised instance segmentation (e.g., BoxInst) is independent, that is, the foreground of an instance is merely distinguished from the background including all the other instances. Under the circumstance of weak supervision, the semantic ambiguity between the neighboring instances tends to be magnified. Independent generation shows the semantic ambiguity in the form of the misclassified pixels in the overlapping area, as illustrated in Fig. 1. Our idea is to simply reduce the overlapping area as small as possible. Speaking of overlapping, we then define an ambiguity measure of overlapping between masks $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$, which is formulated as:

$$overlap_{degree} = \frac{\|\boldsymbol{m}_i \cdot \boldsymbol{m}_j\|_{m_1}}{\|\boldsymbol{m}_i + \boldsymbol{m}_j\|_{m_1}} \tag{1}$$

where $\boldsymbol{m}_i \cdot \boldsymbol{m}_j$ and $\boldsymbol{m}_i + \boldsymbol{m}_j$ can be approximated as the intersection and union between $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$, and $\|\cdot\|_{m_1}$ represents m_1-norm of a matrix.

Next, let's delve into the detailed generation of instance masks. Generally, assuming there are $N+1$ predicted masks, they usually belong to $K+1$ instances, denoted as $M = \{\boldsymbol{m}_0^0, \boldsymbol{m}_1^0 \cdots \boldsymbol{m}_N^K\}$, where $\boldsymbol{m}_n^k$ represents the n-th mask belonging to instance k. We randomly select one of predicted masks from the same instance to form a new set $M_S = \{\boldsymbol{m}^0, \boldsymbol{m}^1 \cdots \boldsymbol{m}^K\}$, where $\boldsymbol{m}^k$ represents the selected mask for the k-th instance. With M and M_S, we apply the ambiguity measure defined in Eq. (1) between the k-th mask in M_S and each mask in M_S not from the k-th instance, and the proposed loss can be formulated as follows:

$$L_{overlap} = \frac{1}{NK}\sum_{n=0}^{N}\sum_{k=0}^{K}\frac{\|\boldsymbol{m}_n^p \boldsymbol{m}^q\|_{m_1}}{\|\boldsymbol{m}_n^p + \boldsymbol{m}^q\|_{m_1}} \quad (p \neq q) \tag{2}$$

where $\boldsymbol{m}_n^p \in M$, and $\boldsymbol{m}^q \in M_S$.

Obviously, our loss promotes mutual exclusiveness between predicted masks of different instances, thus suppressing the generation of overlapping area.

With the projection loss L_{proj} and affinity loss $L_{pairwise}$ in [6], our final loss is formulated as:

$$L = \lambda L_{overlap} + L_{proj} + L_{pairwise} \tag{3}$$

Table 1 . Experimental results on CIS.

Backbone	Overlap Loss	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50-FPN		29.20	64.07	22.87	0.01	4.44	37.68
ResNet-50-FPN	✓	32.62	68.26	27.80	0	6.44	41.55
ResNet-101-FPN		29.39	65.99	22.91	0	4.53	38.00
ResNet-101-FPN	✓	33.09	69.94	27.92	0.02	6.71	42.16
ResNet-101-BiFPN		29.52	65.74	22.69	0.03	5.07	38.12
ResNet-101-BiFPN	✓	32.17	68.89	26.40	0.04	6.97	41.56

where λ is the hyperparameter to adjust the importance of the OverlapLoss.

2.3. CIS Dataset

Crowded scenes are very common in daily life, like party, meeting, sports games, etc. In stark contrast, there is no typical dataset for crowded instance segmentation task, which impedes the development herein. To overcome this, we propose to construct CrowdHuman Instance Segmentation (CIS) dataset.

As is known, CrowdHuman [17] is a popular dataset collected from the Internet for pedestrian detection in crowd, which contains 15,000, 4,370 and 5,000 images for training, validation and testing, respectively. Box annotation in CrowdHuman suits the training of WSIS methods well. For evaluation, we follow the annotation rules in COCO instance segmentation, and have labeled 463 images from Crowdhuman validation dataset, each of which has 3 to 10 people with occlusion. In total, the number of annotated instances reaches 3,453, rendering CIS a convincing dataset. As shown in Fig. 3, the mask annotation is very accurate in crowed scenes.

Fig. 3. Samples from CIS

3. EXPERIMENTS

We evaluate OverlapLoss in both crowded and uncrowded scenes with BoxInst as the base framework. For the former scenario, we train the model on CrowdHuman train split (15,000 images) and evaluate it on CIS (463 images). For the latter, it is trained with train2017 (11,8287 images) and evaluated on test-dev split (20,288 images) from COCO.

3.1. Implementation Details

For the experiments conducted on CrowdHuman and CIS, the batch size is set to 4 (2 images per GPU). Considering dense population in Crowdhuman and CIS but limited GPU memory resources, the training images are randomly resized to have their shorter sides in [384,480] and their longer sides less or equal to 1333. When testing, the scale of shorter side are fixed to 480. According to the ablation study, we set the overlap loss weight λ to 2. Since training with the overlap loss on the first few iterations might fail to converge, all of the experimental results are based on fine-tuning. To be specific, we train the model with 3x (270,000 iterations) learning scheduler to obtain the base model at first and then use the weights to fine-tune with or without overlap loss 1x (90,000 iterations) for comparison. The performance is evaluated with the COCO style mask AP.

Following BoxInst [6], we also conduct experiments on COCO with 16 images per batch (2 images per GPU). For the uncrowded scenes like COCO, we impose a lighter overlapping constraints on the network, thereby setting λ to 1. In this experiment, we further fine-tune 10,000 iterations with learning rate 10^{-4} to get the final results. For a fair comparison, we also fine-tune the baseline 10,000 iterations with the same learning scheduler.

Table 2 . Experimental results on COCO-2017-test-dev.

Backbone	Overlap Loss	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50-FPN		32.2	55.3	32.5	16.2	34.7	42.9
ResNet-50-FPN	✓	32.6	55.6	33.3	16.7	35.1	43.0
ResNet-101-FPN		33.1	56.5	33.6	16.2	35.2	45.1
ResNet-101-FPN	✓	33.5	56.9	34.2	16.6	35.8	45.3
ResNet-101-BiFPN		33.9	57.6	34.3	16.5	36.0	46.4
ResNet-101-BiFPN	✓	34.2	58.0	34.8	17.2	36.3	46.7
ResNet-DCN-101-BiFPN		35.0	59.3	35.6	17.1	37.2	48.6
ResNet-DCN-101-BiFPN	✓	35.4	59.6	36.1	17.6	37.5	49.1

3.2. Quantitive Results

Table 1 lists the results on CIS dataset. It can be seen that the proposed OverlapLoss shows a consistent and striking increase with different backbones in crowded scenes. To be specific, an increase with at least 3.42% 3.7% and 2.65% in AP are obtained with ResNet-50-FPN, ResNet-101-FPN, and ResNet-101-BiFPN respectively. It's self-evident that OverlapLoss effectively ameliorates the semantic ambiguity in crowed scenes and improve the AP for WSIS. Note that when the resolution of testing images increases, a consistent improvement with our loss can be observed. For example, when the shorter edge rescales from 480 to 680, the performance using a ResNet-50-FPN backbone can reach 34.25%.

Table 2 shows the experimental results on uncrowded COCO. It also indicates the effectiveness of OverlapLoss. Though instance occlusion occurs less frequently in COCO, we still observe at least 0.3% increase in terms of AP in all set-ups.

3.3. Qualitative Results

As illustrated in Fig. 4, we present visualization results with and without OverlapLoss on CIS and COCO. Compared with BoxInst [6], the proposed OverlapLoss can produce more accurate segmentation masks especially in heavily crowded scenes, like CIS dataset. It also verifies that our OverlapLoss can deal with the misclassified pixels in occluded areas between instances.

4. CONCLUSIONS

In the paper, we discuss the forthcoming challenges in crowded scenes for weakly supervised instance segmentation, that is, the independent mask inference and the lack of evaluation dataset. We thus propose a new loss of OverlapLoss which can be appended to any off-the-shelf framework with ease and gains decent increase in both crowded and uncrowded scenes. Besides, to facilitate the research in crowded scenes, we also provide a new evaluation dataset.

Fig. 4. Visualization of instance segmentation on CIS and COCO. As shown in red dash boxes the proposed OverlapLoss can generate higher quality instance segmentation.

5. REFERENCES

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2209–2218.

[2] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao, "Weakly supervised instance segmentation using class peak response," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3791–3800.

[3] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov, "Pointly-supervised instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2617–2626.

[4] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xian-Sheng Hua, and Lei Zhang, "Box-supervised instance segmentation with level set evolution," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*. Springer, 2022, pp. 1–18.

[5] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar, "Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3406–3416.

[6] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen, "Boxinst: High-performance instance segmentation with box annotations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5443–5452.

[7] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Qian Zhang, and Wenyu Liu, "Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation," *arXiv preprint arXiv:2210.05174*, 2022.

[8] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang, "Weakly supervised instance segmentation using the bounding box tightness prior," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] Zhi Tian, Chunhua Shen, and Hao Chen, "Conditional convolutions for instance segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 282–298.

[10] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo, "Polarmask: Single shot instance segmentation with polar representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12193–12202.

[11] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9157–9166.

[12] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng, "End-to-end people detection in crowded scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2325–2333.

[13] Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun, "Detection in crowded scenes: One proposal, multiple predictions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12214–12223.

[14] Zixuan Xu, Banghuai Li, Ye Yuan, and Anhong Dang, "Beta r-cnn: Looking into pedestrian detection from another perspective," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19953–19963, 2020.

[15] Jiangfan Deng, Dewen Fan, Xiaosong Qiu, and Feng Zhou, "Improving crowded object detection via copy-paste," *arXiv preprint arXiv:2211.12110*, 2022.

[16] Maria Lyssenko, Christoph Gladisch, Christian Heinzemann, Matthias Woehrle, and Rudolph Triebel, "Instance segmentation in carla: Methodology and analysis for pedestrian-oriented synthetic data generation in crowded scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 988–996.

[17] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun, "Crowdhuman: A benchmark for detecting human in a crowd," *arXiv preprint arXiv:1805.00123*, 2018.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.