(a) (b) (c) (d) (e) (f)

Figure A: The failure examples produced by our proposed method.

We thank reviewers for their comments which are very helpful for improving the paper. We address reviewers' comments in the rebuttal and will revise the paper accordingly. All the results are evaluated by $\mathrm{mAP}^r_{0.5}$ unless specified otherwise.

R1: *About the tightness assumption. How sensitive/robust is the proposed method?* As suggested, we evaluate the proposed method by expanding/contracting the bounding boxes and show its performance under different expansion/contraction ratios in following table. The tightness prior is helpful to construct the training bags, and the proposed method is quite robust to the small ratios from -5% to $+5\%$. Moreover, by slightly contracting the bounding boxes, e.g. -5%, we can reduce noisy pixels in each positive bag, resulting in even better performance. However, excessively expanding or contracting ratios leads to unreliable positive and negative bags, and thus produces sub-optimal results.

Ratio	$+15\%$	$+10\%$	$+5\%$	0%	-5%	-10%	-15%
Ours	54.0	55.0	58.0	58.9	59.5	46.0	28.1

R1: *About the CVPR'18 paper "Learning ..."* We will cite it, and discuss its similarities and differences from our paper.

R2: *About average pooling.* Only with the unary term, the result of average pooling is 36.8, which falls behind the result of max pooling, 43.9, because average pooling considers all pixels in the positive bags as the foreground and overestimates the object region. Therefore, using max pooling can better diminish false alarms.

R2: *Result on COCO and small objects.* The result on *coco minival* is shown in following table (BoxMask is the baseline method in the paper). Our method outperforms the baseline and reaches 78.3% $(= 45.5/58.1)$ and 68.3% $(= 11.2/16.4)$ of the performance of fully-supervised Mask R-CNN in $\mathrm{AP}^r_{0.5}$ and AP^r_S, respectively. The results show that our method is effective in segmenting diverse objects with varied sizes.

method	AP	$\mathrm{AP}^r_{0.5}$	$\mathrm{AP}^r_{0.75}$	AP^r_S	AP^r_M	AP^r_L
BoxMask	11.1	31.1	6.0	5.3	11.6	15.9
Ours	21.1	45.5	17.2	11.2	22.0	29.8
Mask R-CNN [7]	36.3	58.1	38.5	16.4	38.9	53.5

R2: *Failure cases.* Figure A shows failure cases of our method. In (a) and (b), segments of small objects are incomplete due to inaccurate boundary on low-resolution regions. In (c) and (d), different instances are wrongly merged. In (e) and (f), inaccurate object contours are segmented due to inter-instance similarity and cluttered scenes.

R3: *Performance versus annotation effort.* Based on [Bellver et al, Budget-aware Semi-Supervised Semantic and Instance Segmentation, CVPR'19 workshop], the instance-level and box-level annotation costs of Pascal VOC are 239.7 and 38.1 seconds per image, respectively. We train Mask R-CNN by instance-level annotation, and limit the amount of annotation so that the annotation budget is comparable with $1.0\times, 1.5\times, 2.0\times$ of the box-level annotation. The results of the three different settings are 48.3, 53.5 and 59.9, respectively. The first two results fall behind our method by the margins 10.6 and 5.4, while the last one surpasses ours by 1.0 but with $2\times$ annotation cost. Therefore, with the same annotation cost, our method outperforms Mask R-CNN because less training data lead to overfitting for Mask R-CNN.

R3: *The four questions about Eq.4.* (1) Epsilon is an 8-neighbor set. (2) Because neighboring pixels are connected in the pairwise term, we can propagate the segment scores, and thus enlarge the segment. (3) We assume "patches" in the review refers to "bags" in our method. If all instances in positive bags are treated as positive samples, many background pixels are mistaken as positive samples. Therefore, more false alarms could be predicted. (4) With the tightness prior, each positive bag meets the MIL assumption, so this task could be solved with the MIL formulation.

R3: *Comparing efficiency with [16].* The method [16] requires pre-generated proposals from Selective Search, obtains the detected boxes by Fast R-CNN, and finally takes the detected boxes and the RGB images as the input to generate the instance result. In [16], the proposal generation step alone takes ≈ 10 seconds per image. In contrast, our method requires only one forward pass which takes only ≤ 0.1 seconds per image or ≈ 1 second per image if DenseCRF is applied. Therefore, our method is more efficient than [16].

R3: *About L85.* We agree with this comment, and will make it more clear in the revision. However, what we want to claim is that in general, the applicability of the fully supervised methods "may" be limited in the real world because of the high annotation cost.

R3: *Comparing performance with fully supervised Mask-RCNN.* Our method adopting the MIL formulation tends to highlight the discriminative parts of objects, while Mask R-CNN with mask-level annotation emphasizes the whole objects. The IOUs between the ground truth and the discriminative regions are often larger than 0.25 but less than 0.5. It is why our method slightly outperforms Mask R-CNN in $\mathrm{mAP}^r_{0.25}$, but falls behind it in $\mathrm{mAP}^r_{0.5}$, $\mathrm{mAP}^r_{0.7}$, and $\mathrm{mAP}^r_{0.75}$, as reported in Table 1 of the submitted manuscript.