Mixed-dual-head Meets Box Priors: A Robust Framework for Semi-supervised Segmentation – Supplementary Material –

Chenshu Chen Tao Liu Wenming Tan Shiliang Pu
Hikvision Research Institute
{chenchenshu, liutao46, tanwenming, pushiliang.hri}@hikvision.com

A. Pseudo code of training

We summary one training round of our framework in Algorithm 1. The training set consists of two parts: the fully labeled dataset $\mathcal{F} = \{(x_i, y_i, b_i)\}_{i=1}^{F}$ containing F fully labeled images and the weakly labeled dataset $\mathcal{W} = \{(x_i, b_i)\}_{i=1}^{W}$ containing W weakly labeled images, where x_i denotes i-th image, y_i and b_i denote its ground-truth mask and bounding box label.

B. The effectiveness of priors

We show the additional experiment results about adding priors or not in Table 9. It can be seen that both position and class priors can improve the performance, and combining the both can yield further improvements.

	Backbone	F	W	mIoU
w/o priors	ResNet-50	1464	9118	79.30
w/ position priors	ResNet-50	1464	9118	79.78
w/ class priors	ResNet-50	1464	9118	80.49
w/ both priors	ResNet-50	1464	9118	81.31
w/o priors	ResNet-101	1464	9118	80.65
w/ position priors	ResNet-101	1464	9118	81.14
w/ class priors	ResNet-101	1464	9118	82.56
w/ both priors	ResNet-101	1464	9118	82.82

Table 9. Additional experiment results for Table 2 of the main paper on PASCAL VOC 2012 val set. F and W are the numbers of fully labeled images and weakly labeled images respectively.

C. Selection of AnnNet

Table 10 shows the performance of SegNet when choosing different network as AnnNet. It can be seen that SegNet can achieve better results with a stronger AnnNet.

Algorithm 1: One Training Round of Our Framework

Input: Fully labeled dataset $\mathcal{F} = \{(x_i, y_i, b_i)\}_{i=1}^{F}$, weakly labeled dataset $\mathcal{W} = \{(x_i, b_i)\}_{i=1}^{W}$, a mixing ratio r, AnnNet g_φ parameterized by φ, SegNet f_θ parameterized by θ

Output: SegNet f_θ

1. AnnNet training:

Obtain proposal masks $\mathcal{Y}^p = \{y_i^p\}_{i=1}^{W}$ for $\mathcal{W}$ as in Figure 2 of the main paper.

Update $\mathcal{W} = \{x_i, y_i^p, b_i\}_{i=1}^{W}$.

for *each iteration* **do**

- Sample a batch $\mathcal{B}^S$ from $\mathcal{F}$ and $\mathcal{W}$ with a mixing ratio r for strong head.
- Sample a batch $\mathcal{B}^W$ from $\mathcal{W}$ for weak head.
- Add position and class priors for images in $\mathcal{B}^S$ and $\mathcal{B}^W$.
- Perform a forward inference on $\mathcal{B}^S$ and $\mathcal{B}^W$ using g_φ.
- Calculate loss and update φ by SGD.

end

2. Pseudo mask generation:

Obtain prediction results $\hat{\mathcal{Y}} = \{\hat{y}_i\}_{i=1}^{W}$ for $\mathcal{W}$ using g_φ.

Obtain pseudo masks $\tilde{\mathcal{Y}} = \{\tilde{y}_i\}_{i=1}^{W}$ by Eq.(3).

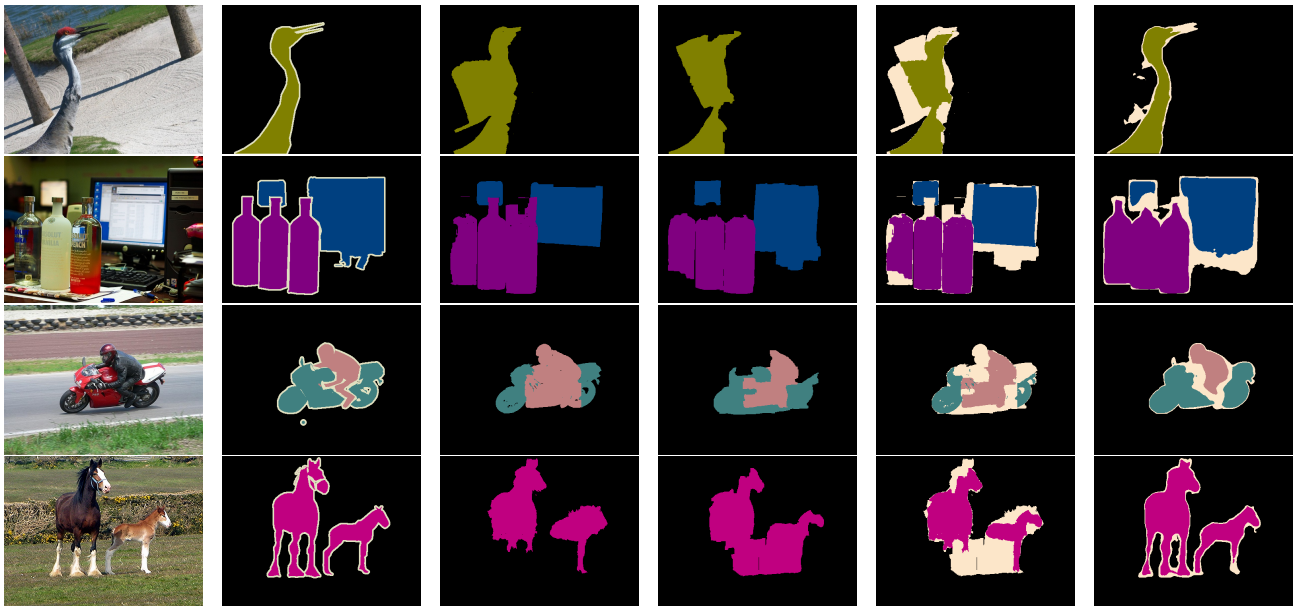
Update $\mathcal{W} = \{x_i, \tilde{y}_i, b_i\}_{i=1}^{W}$.

3. SegNet training:

for *each iteration* **do**

- Sample a batch $\mathcal{B}^S$ from $\mathcal{F}$ and $\mathcal{W}$ with a mixing ratio r for strong head.
- Sample a batch $\mathcal{B}^W$ from $\mathcal{W}$ for weak head.
- Perform a forward inference on $\mathcal{B}^S$ and $\mathcal{B}^W$ using f_θ.
- Calculate loss and update θ by SGD.

end



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

Figure 9. Visualization of proposal masks and pseudo masks. (a) Image. (b) Ground truth. (c) Proposal masks produced by GrabCut. (d) Proposal masks produced by MCG. (e) Proposal masks by fusing (c) and (d). (f) Pseudo masks produced by AnnNet.

AnnNet	SegNet	F	W	MS	mIoU
Deeplabv3+ ResNet-101	Deeplabv3+ ResNet-50	1464	9118		80.20
Deeplabv3+ ResNet-101	Deeplabv3+ ResNet-50	1464	9118	✓	81.30
Deeplabv3+ ResNet-101	Deeplabv3+ ResNet-101	1464	9118		81.52
Deeplabv3+ ResNet-101	Deeplabv3+ ResNet-101	1464	9118	✓	83.19
HRNetV2-W48	Deeplabv3+ ResNet-50	1464	9118		81.31
HRNetV2-W48	Deeplabv3+ ResNet-50	1464	9118	✓	82.26
HRNetV2-W48	Deeplabv3+ ResNet-101	1464	9118		82.82
HRNetV2-W48	Deeplabv3+ ResNet-101	1464	9118	✓	83.78

Table 10. Results on PASCAL VOC 2012 val set when using different networks for AnnNet and SegNet. "MS" denotes using multi-scale and left-right flipped inputs at inference time.

(a) (b) (c) (d)

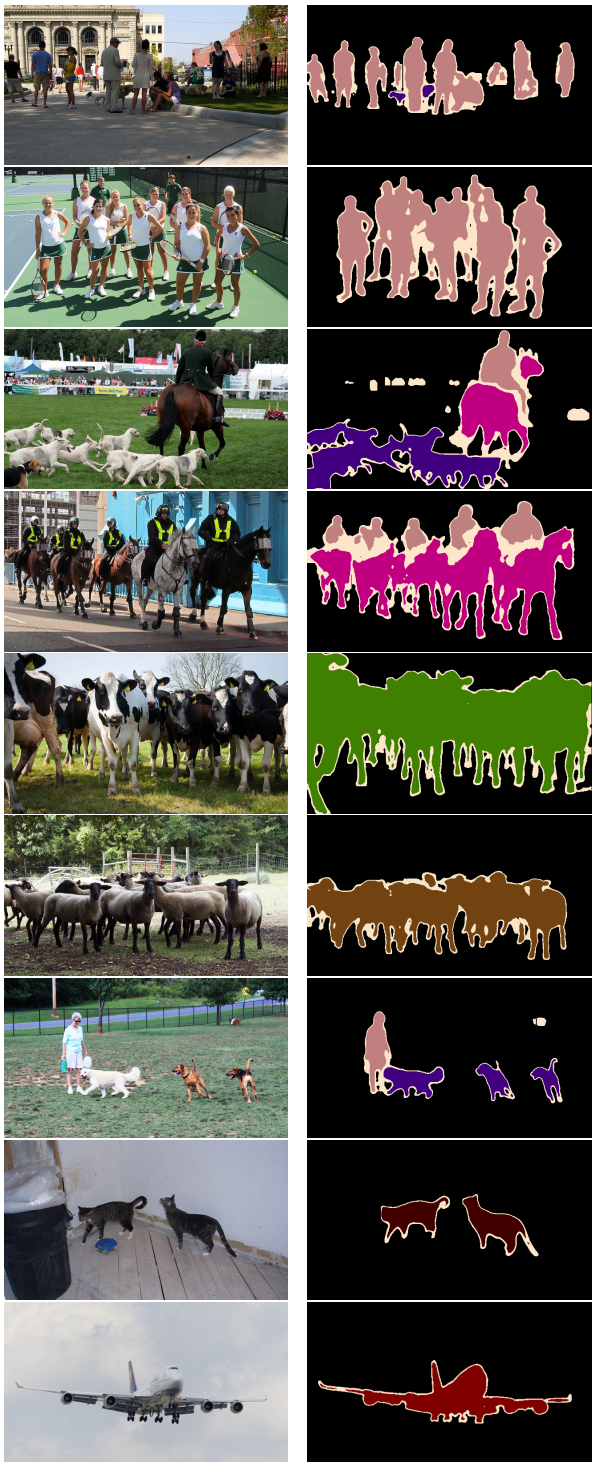
Figure 8. Visualization of pseudo masks produced by AnnNet in different rounds of iterative training. (a) Image. (b) Ground truth. (c) Pseudo masks in round 1. (d) Pseudo masks in round 4.

D. Qualitative results of iterative training

Figure 8 shows some pseudo masks generated by the AnnNet in different rounds of iterative training. It shows that the pseudo masks can be further refined through iterative training.

E. Segmentation labels

Figure 9 shows different segmentation labels. It can be seen that the original proposal masks produced by GrabCut (Figure 9(c)) and MCG (Figure 9(d)) are very coarse. Though the fused results (Figure 9(e)) of GrabCut and MCG proposals can ignore some potential mislabeled pixels, they are still far from precise. In comparison, the pseudo masks generated by the AnnNet (Figure 9(f)) are much more accurate.



(a) (b)

Figure 10. Visualization of some pseudo masks of COCO dataset. The first six rows show some complex scenarios, and the last three rows show some simple scenarios. (a) Image. (b) Pseudo mask

F. Pseudo masks of COCO dataset

Figure 10 shows some pseudo masks of COCO dataset generated by the AnnNet. It can be seen that AnnNet is able to generate high quality pseudo masks for simple scenarios. Even in more complex scenarios, the pseudo masks are also satisfying.