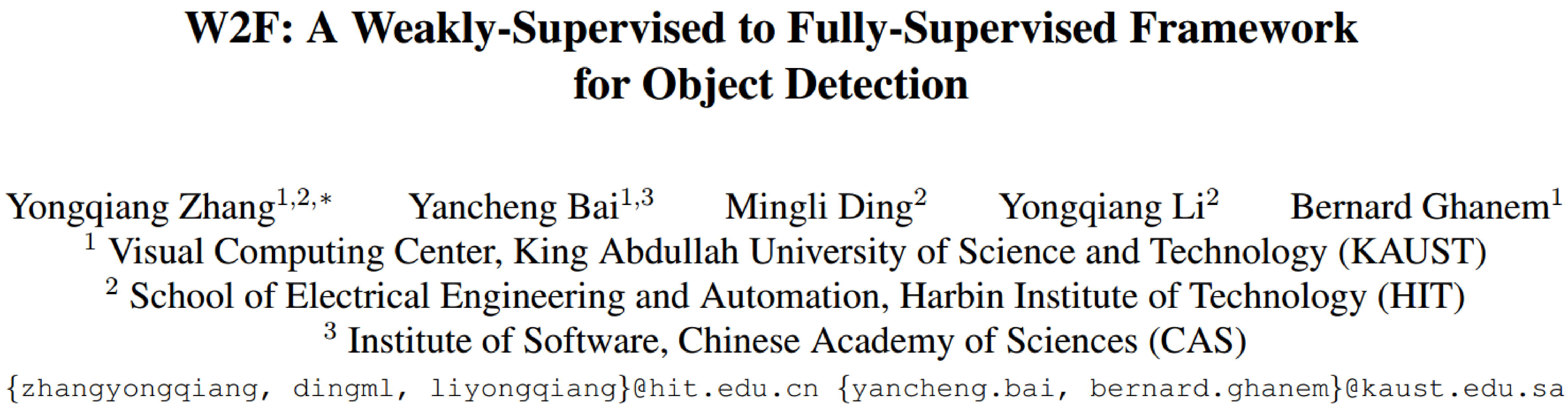


W2F: A Weakly-Supervised to Fully-Supervised Framework for Object Detection

Yongqiang Zhang[1,2,*] Yancheng Bai[1,3] Mingli Ding[2] Yongqiang Li[2] Bernard Ghanem[1]
[1] Visual Computing Center, King Abdullah University of Science and Technology (KAUST)
[2] School of Electrical Engineering and Automation, Harbin Institute of Technology (HIT)
[3] Institute of Software, Chinese Academy of Sciences (CAS)
{zhangyongqiang, dingml, liyongqiang}@hit.edu.cn {yancheng.bai, bernard.ghanem}@kaust.edu.sa

Hengtao Guo

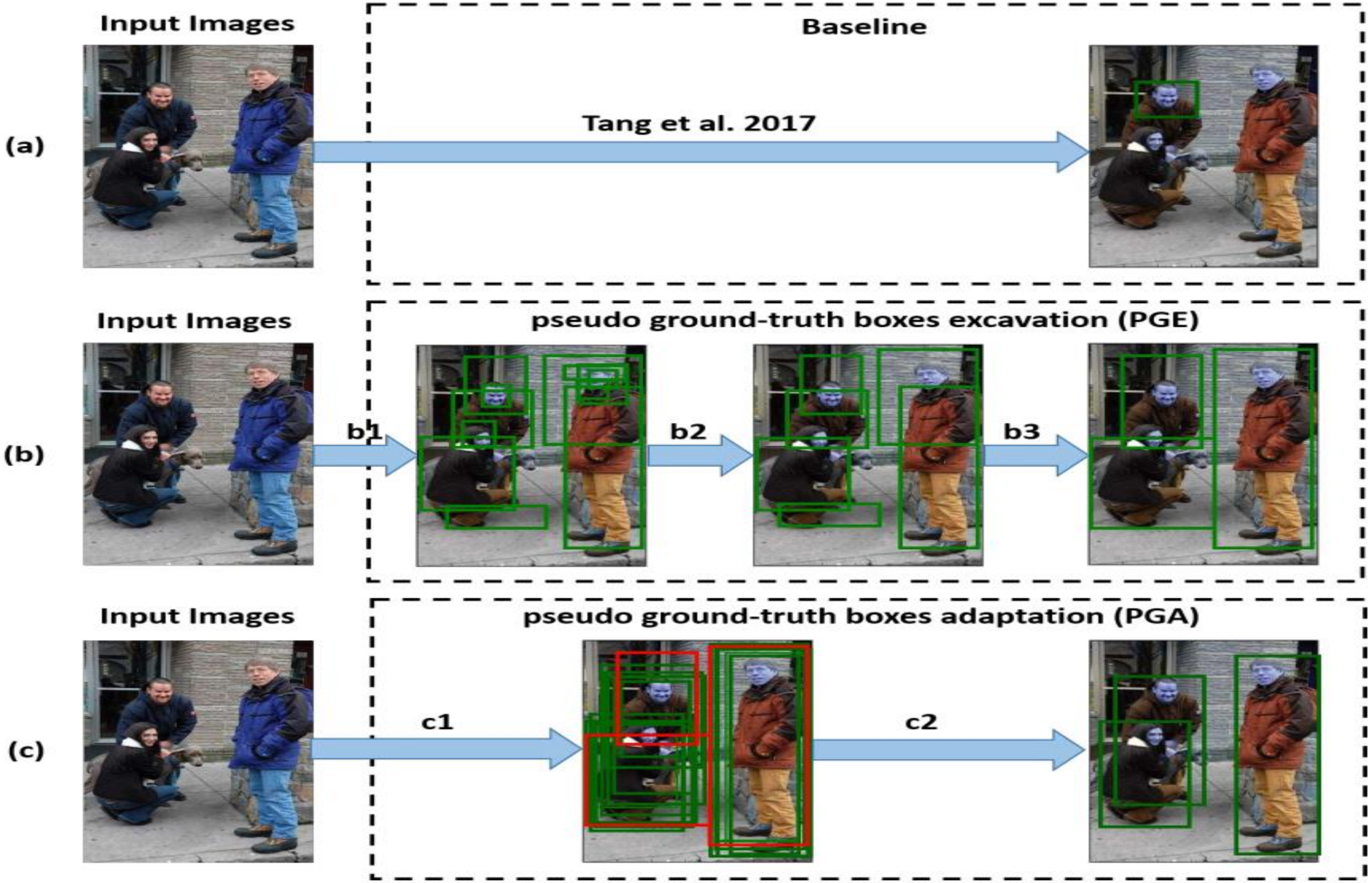
Main ideas

1. Weakly-supervised object detection: data is enough, but accuracy is limited
2. Fully-supervised object detection: performance is good, but annotation is troublesome
3. **Target**: Weakly-supervised to fully-supervised framework

4-step framework

1. A weakly-supervised detector (**WSD**) is implemented using multiple instance learning.
2. A pseudo ground-truth excavation (**PGE**) algorithm to find the pseudo ground-truth of each instance in the image.
3. The pseudo ground-truth adaptation (**PGA**) algorithm is designed to further refine the pseudo ground-truths from PGE.
4. Use these pseudo ground-truths to train a fully-supervised detector (**FSD**).
5. ***WSD + PGE + PGA + FSD***

Overall framework


Input Images
Baseline
(a)
Tang et al. 2017
Input Images
pseudo ground-truth boxes excavation (PGE)
(b)
b1
b2
b3
Input Images
pseudo ground-truth boxes adaptation (PGA)
(c)
c1
c2

Weakly Supervised Detection

1. Learn the discriminative representation of the object instances
2. Select them from positive images to train a detector.

Cat True
Dog True
Bird False

WSDNN

1. Given an image I, denote the image-level labels

$$\mathbf{y}=[y_1, y_2, \ldots, y_C] \in \mathbb{R}^{n\times 1}$$

2. For each input image I, object proposals

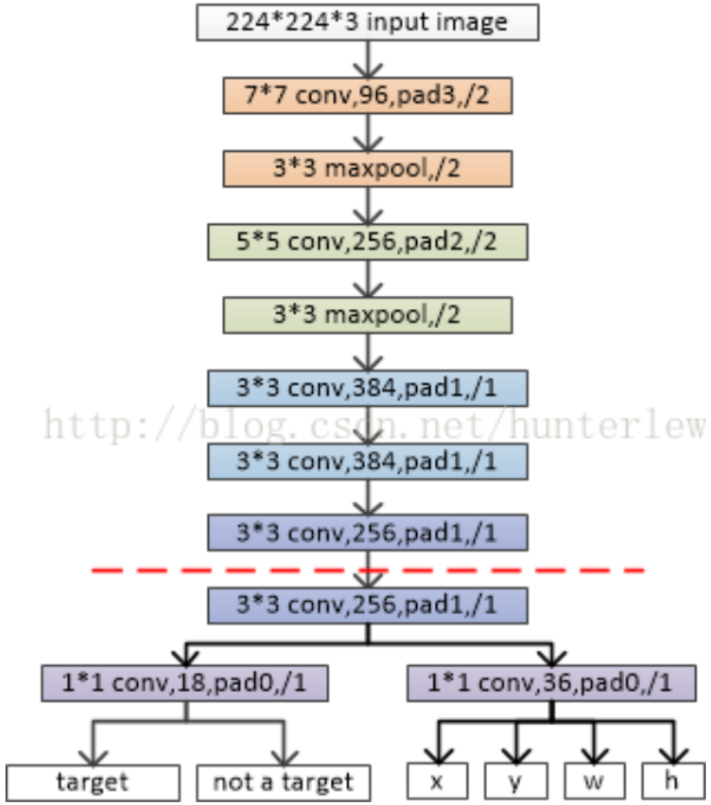
$$\mathbf{R} = (r_1, \ldots, r_n)$$

3. The first stream performs classification

$$[\sigma_{class}(\mathbf{x}^c)]_{ij}=\frac{e^{x^c_{ij}}}{\sum_{k=1}^{C} e^{x^c_{kj}}}$$

4. The second stream performs instead detection

$$[\sigma_{det}(\mathbf{x}^d)]_{ij}=\frac{e^{x^d_{ij}}}{\sum_{k=1}^{|R|} e^{x^d_{ik}}}$$



WSDNN

1. The score of each proposal

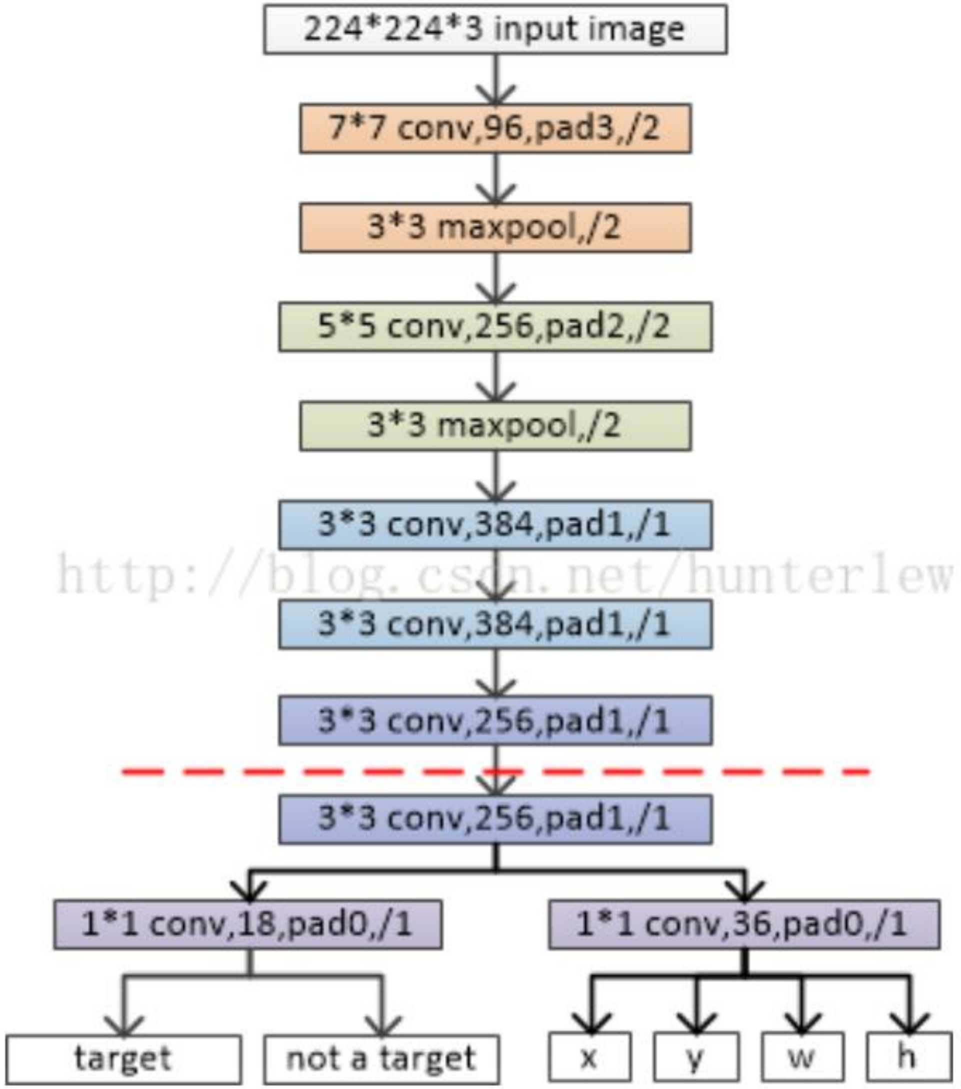
$$\mathrm{x}^{R} = \sigma_{class}(\mathbf{x}^{c}) \odot \sigma_{det}(\mathbf{x}^{d})$$

2. The cth class prediction score at the image-level can be obtained by summation over all proposals:

$$p_c = \sum_{r=1}^{|R|} x_{cr}^{R}.$$

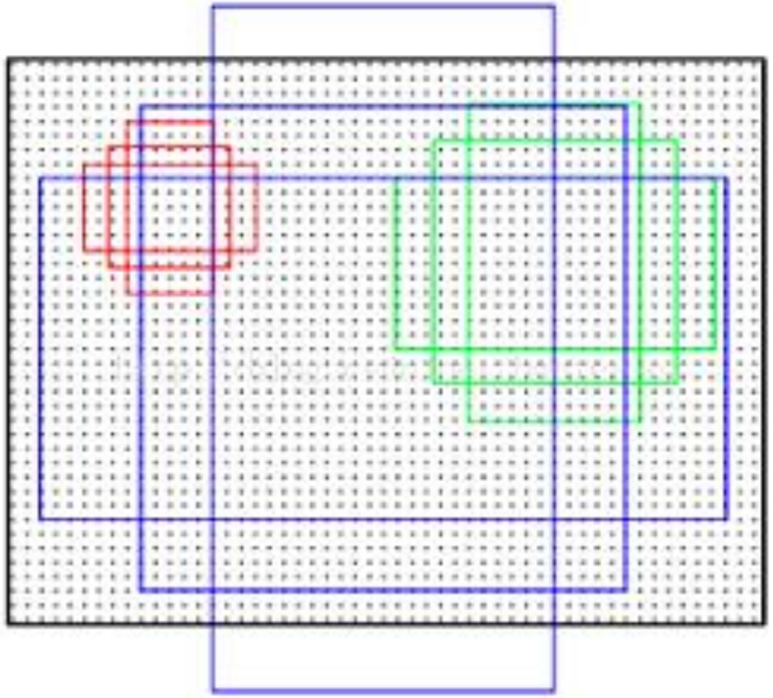
3. Loss function:

$$Loss_w = -\sum_{c=1}^{C}\{y_c \log p_c + (1-y_c)\log(1-p_c)\}$$



Size: 8, 16, 32

Scale: 0.5, 1, 2



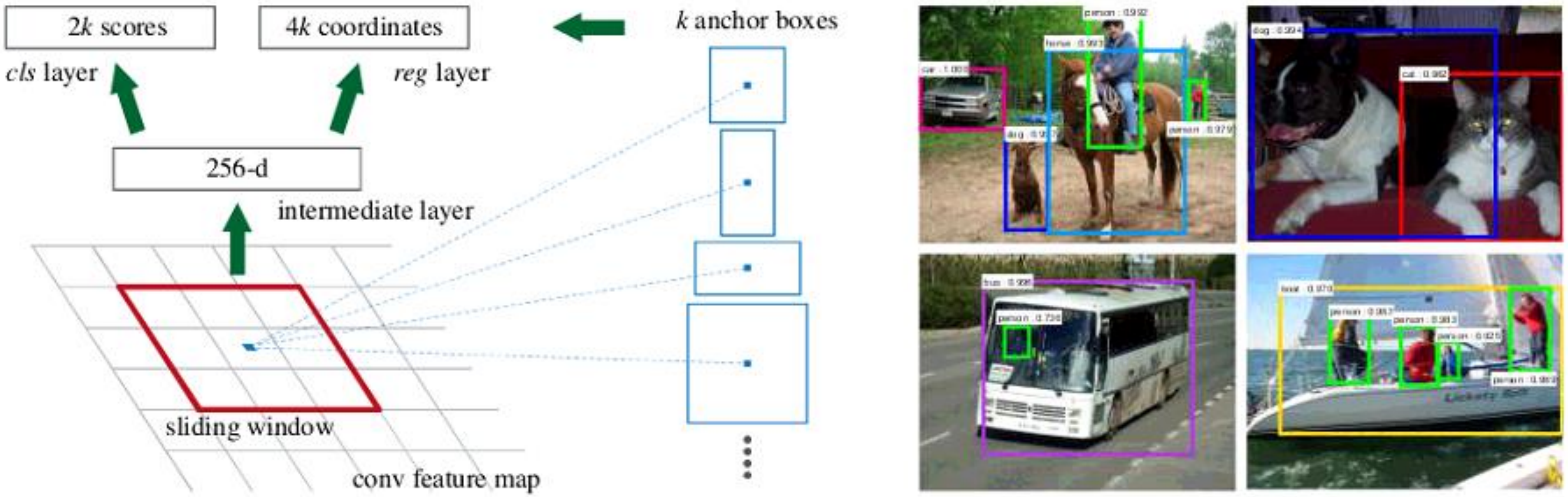
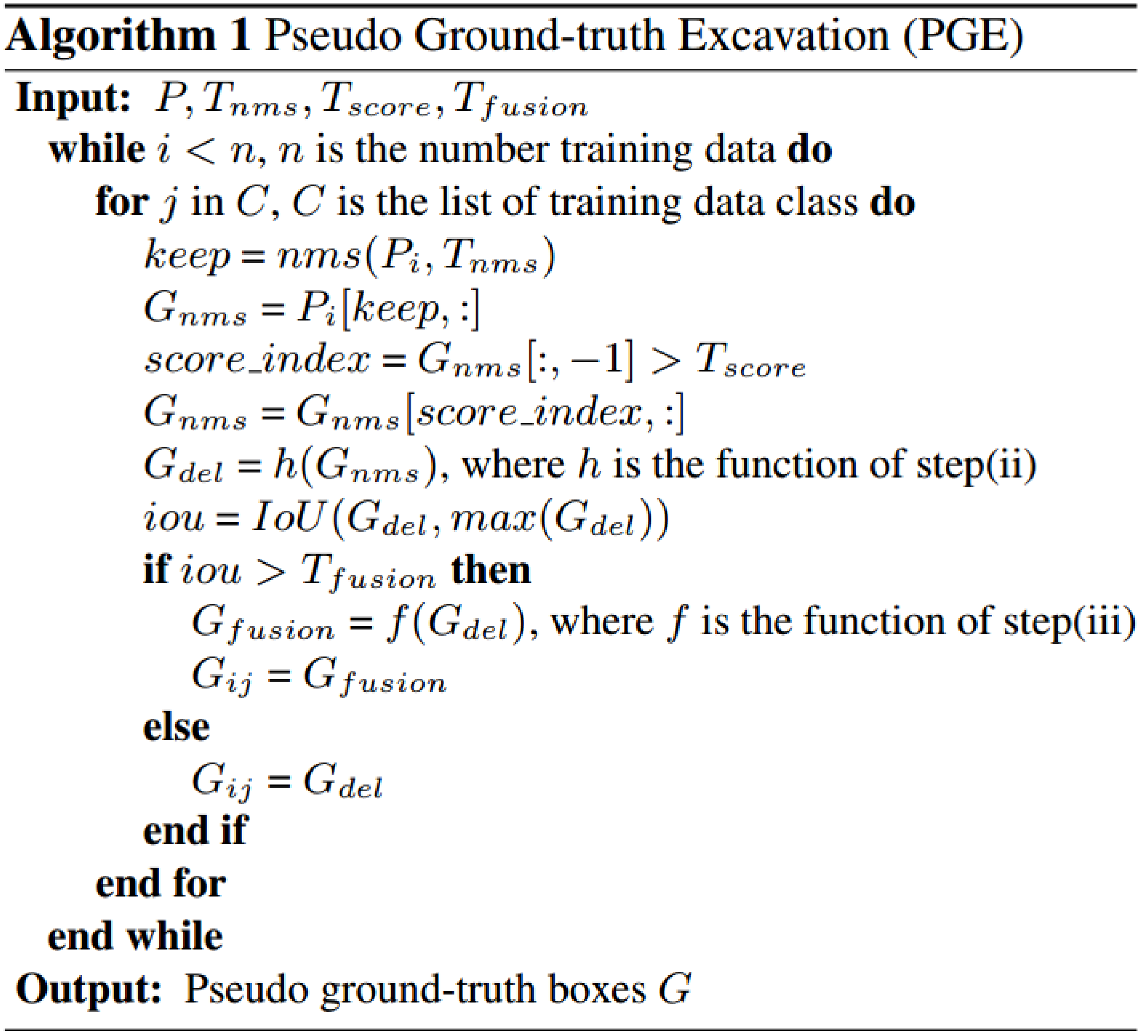


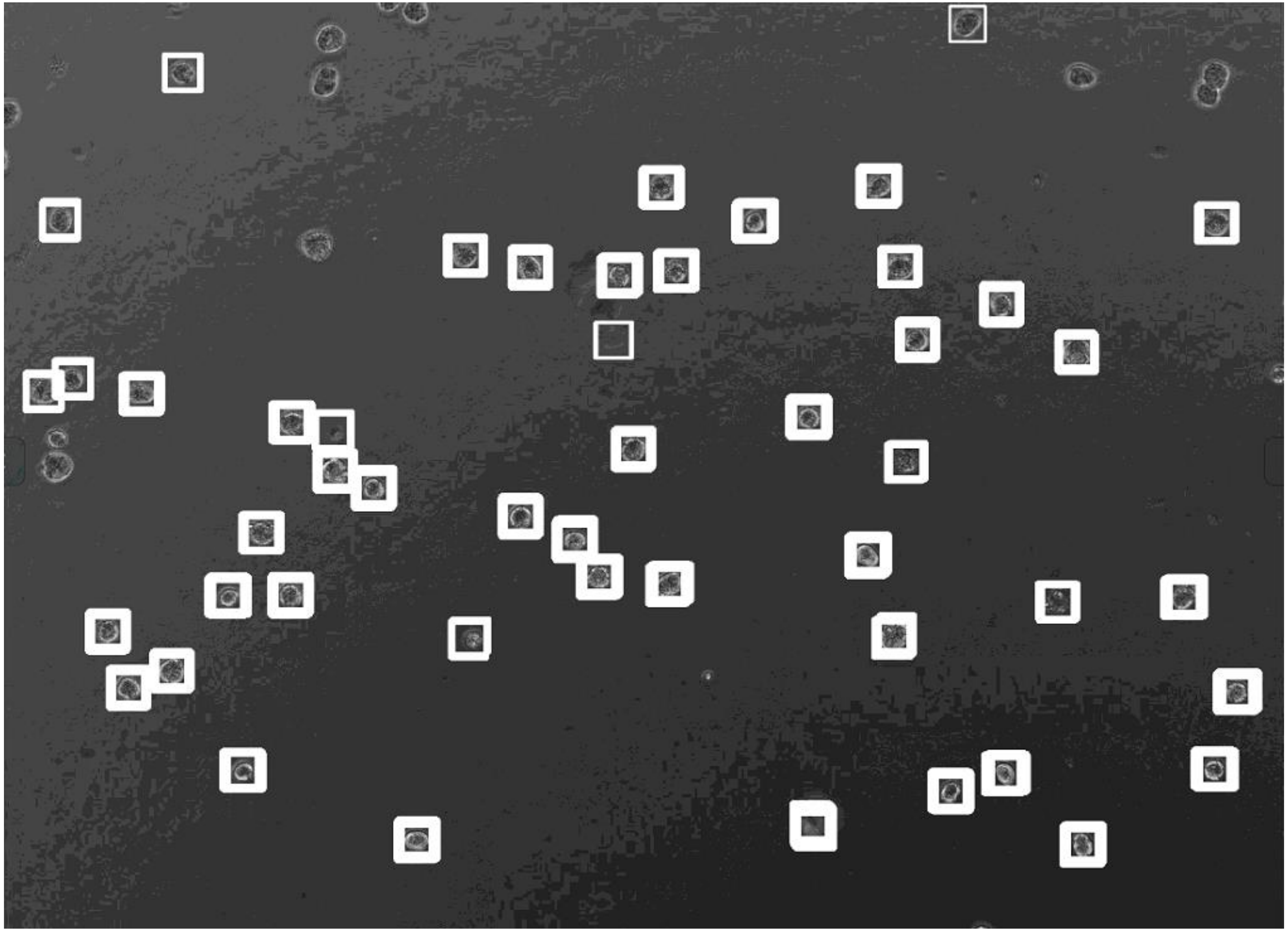
Figure 1: **Left**: Region Proposal Network (RPN). **Right**: Example detections using RPN proposals on PASCAL VOC 2007 test. Our method detects objects in a wide range of scales and aspect ratios.

Pseudo ground-truth excavation (PGE)

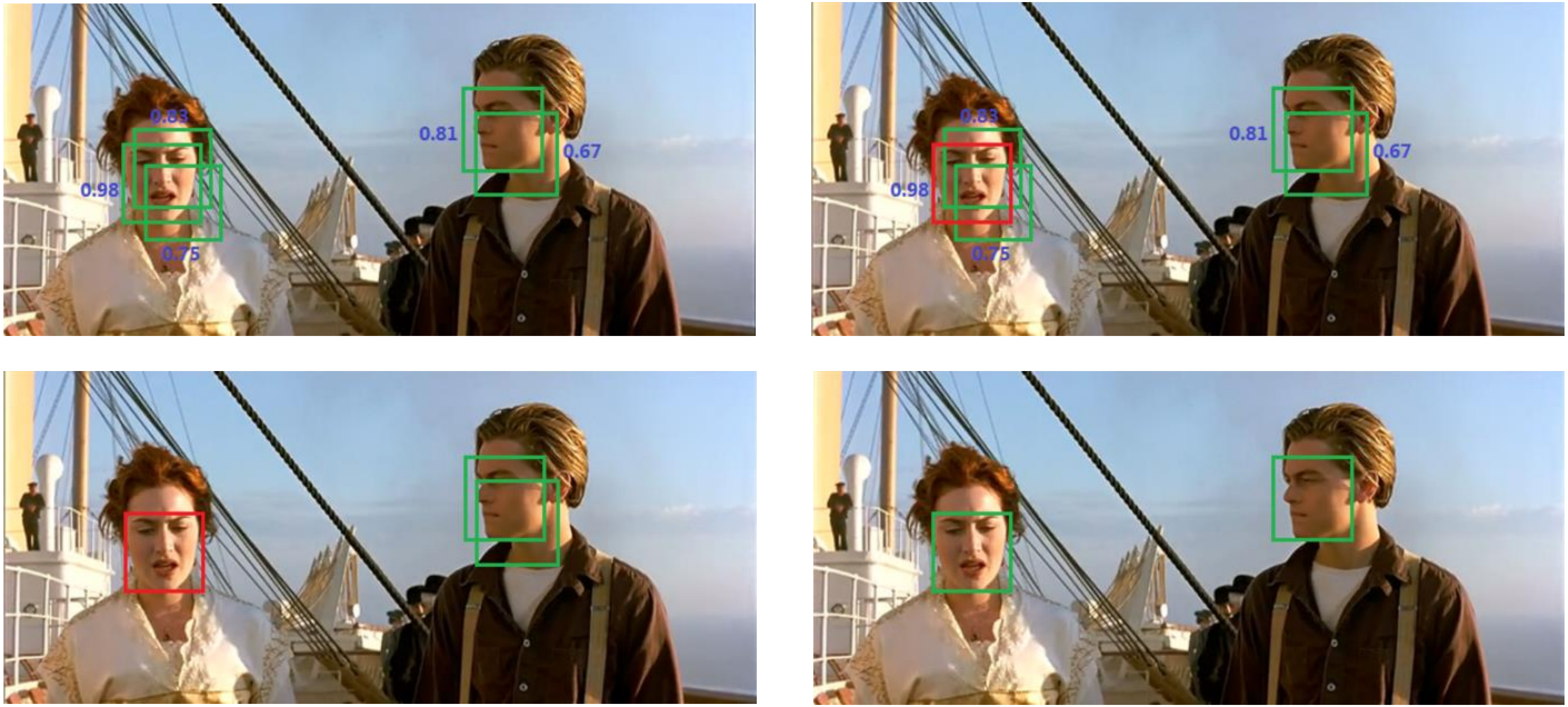


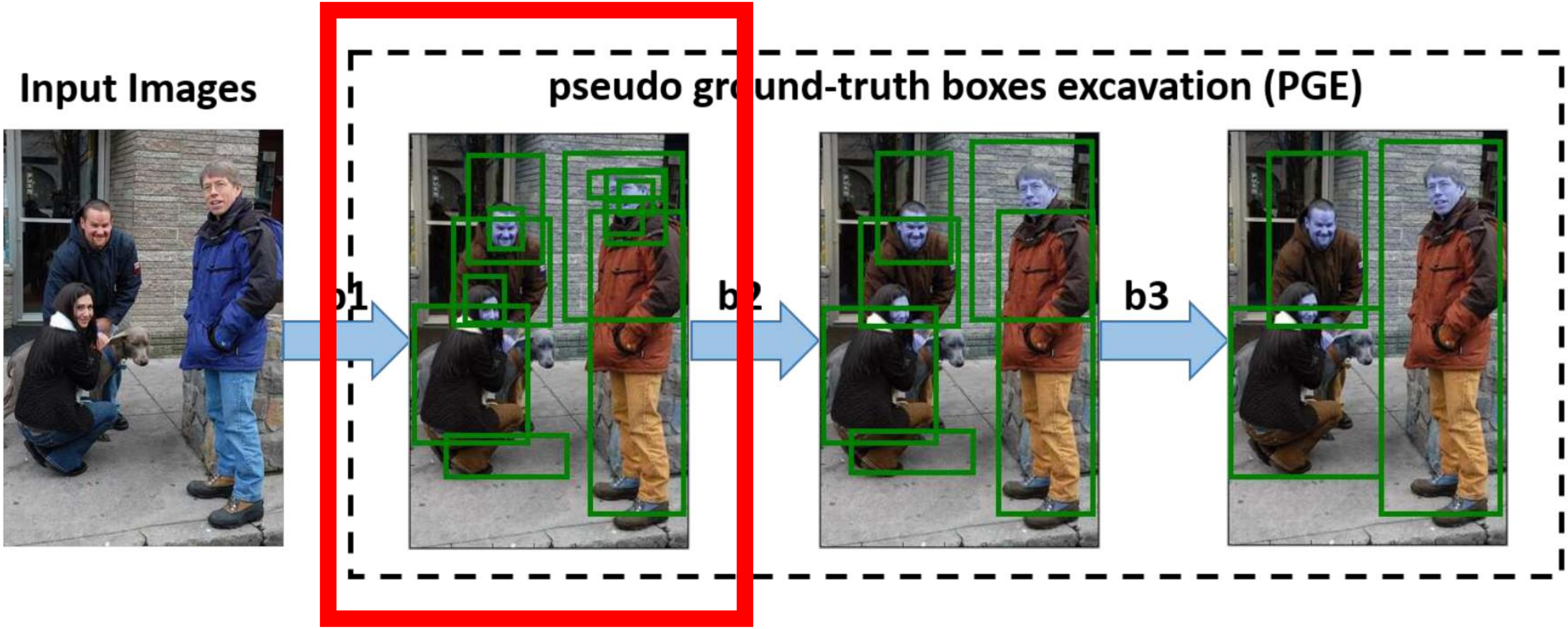
Algorithm 1 Pseudo Ground-truth Excavation (PGE)

```
Input: $P, T_{nms}, T_{score}, T_{fusion}$
  while $i < n$, $n$ is the number training data do
    for $j$ in $C$, $C$ is the list of training data class do
      $keep = nms(P_i, T_{nms})$
      $G_{nms} = P_i[keep, :]$
      $score\_index = G_{nms}[:, -1] > T_{score}$
      $G_{nms} = G_{nms}[score\_index, :]$
      $G_{del} = h(G_{nms})$, where $h$ is the function of step(ii)
      $iou = IoU(G_{del}, max(G_{del}))$
      if $iou > T_{fusion}$ then
        $G_{fusion} = f(G_{del})$, where $f$ is the function of step(iii)
        $G_{ij} = G_{fusion}$
      else
        $G_{ij} = G_{del}$
      end if
    end for
  end while
Output: Pseudo ground-truth boxes $G$
```



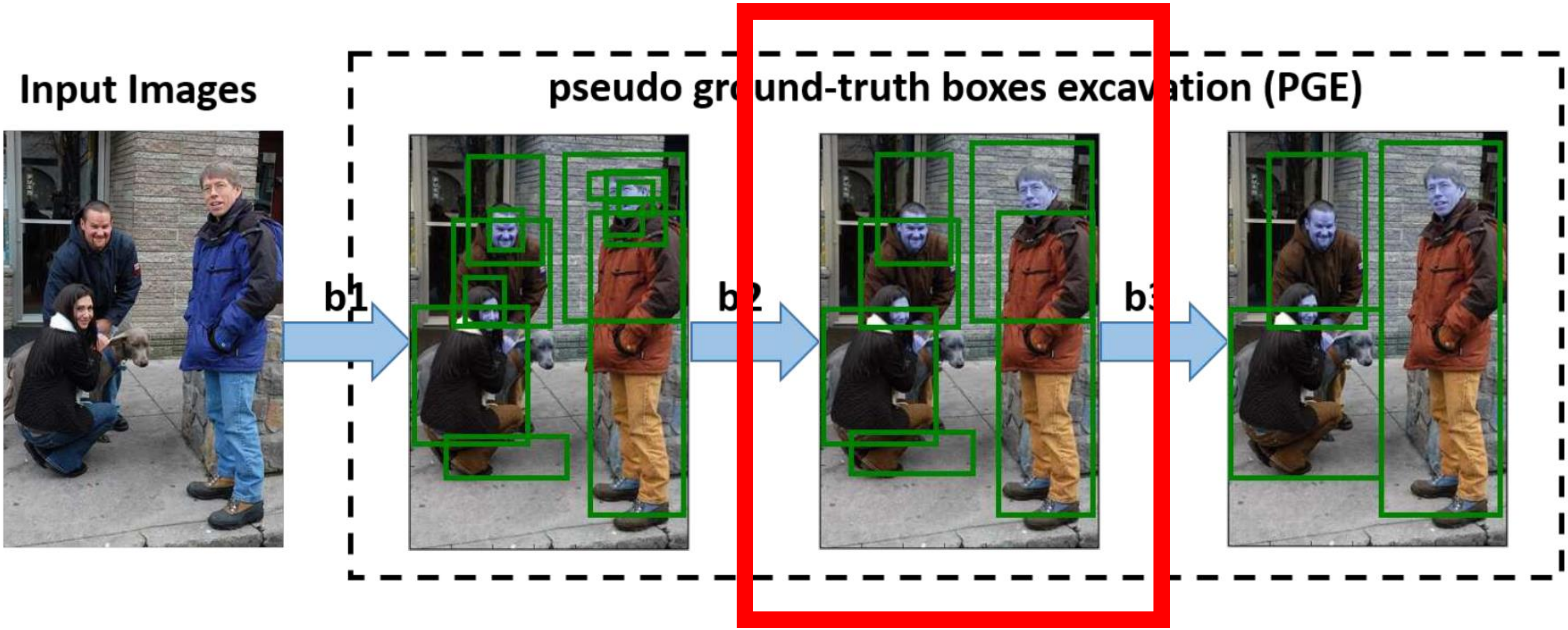
Non maximum suppression





Input Images

pseudo ground-truth boxes excavation (PGE)



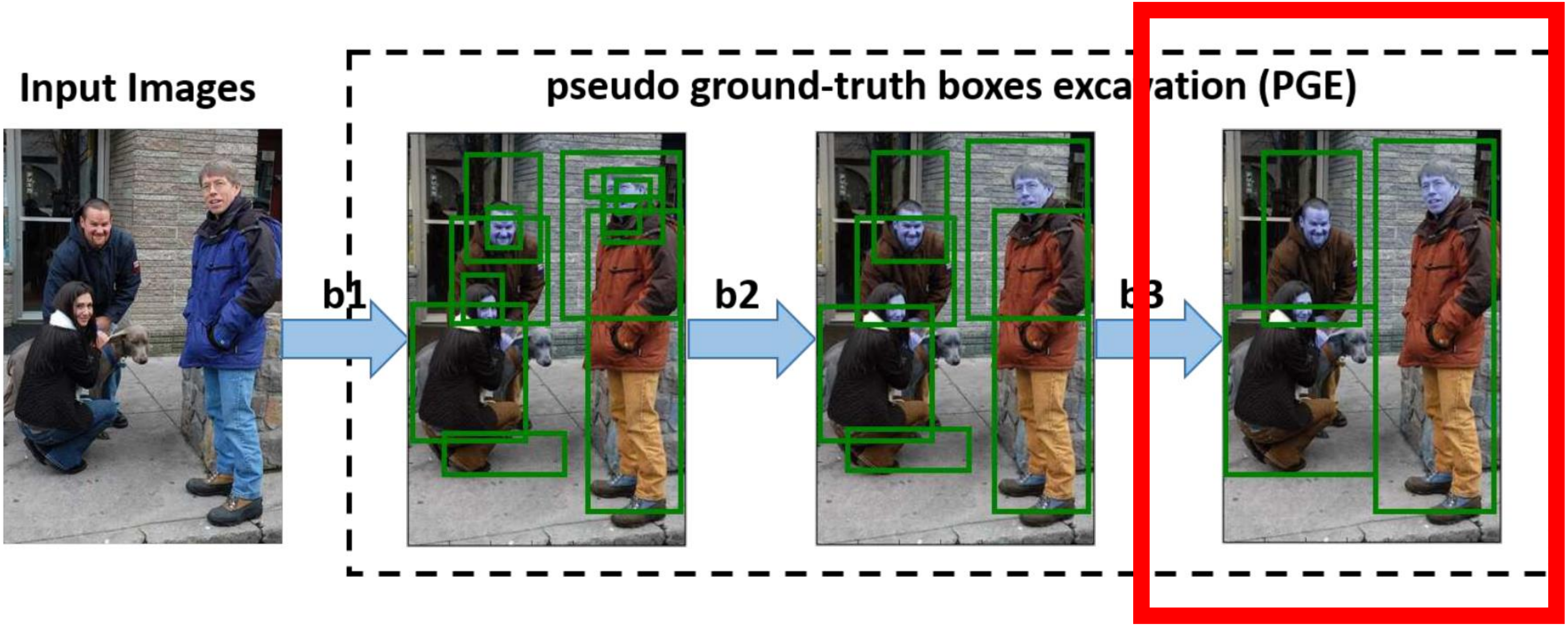
Input Images

pseudo ground-truth boxes excavation (PGE)

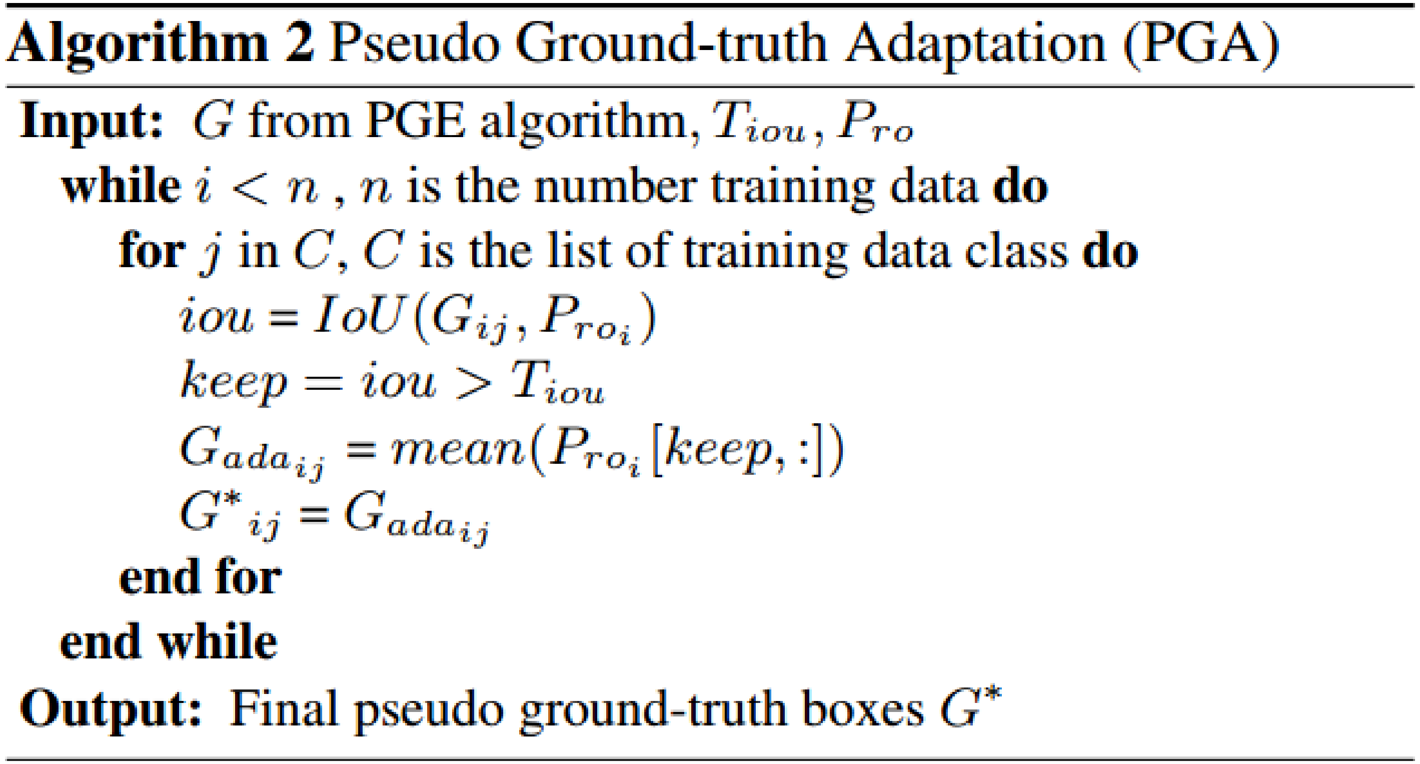
b1

b2

b3

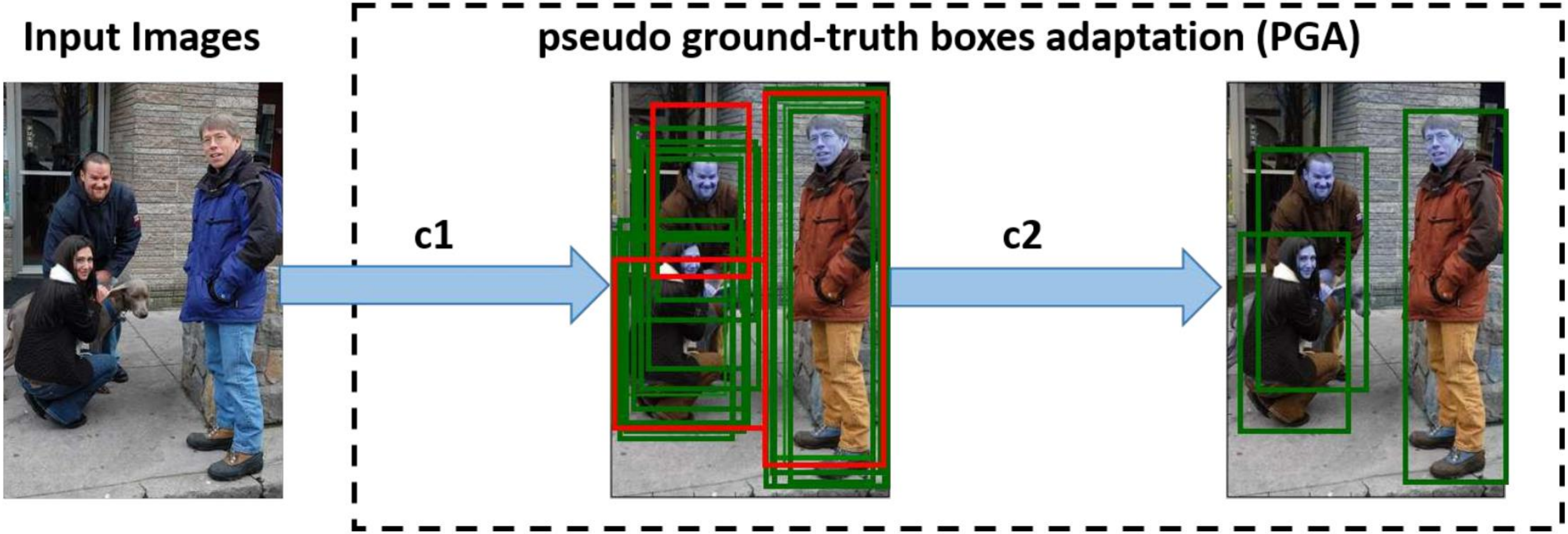

Input Images
pseudo ground-truth boxes excavation (PGE)
b1
b2
b3

Pseudo ground-truth adaptation (PGA)



Algorithm 2 Pseudo Ground-truth Adaptation (PGA)

```
Input:  $G$ from PGE algorithm, $T_{iou}$, $P_{ro}$
  while $i < n$ , $n$ is the number training data do
    for $j$ in $C$, $C$ is the list of training data class do
      $iou = IoU(G_{ij}, P_{ro_i})$
      $keep = iou > T_{iou}$
      $G_{ada_{ij}} = mean(P_{ro_i}[keep, :])$
      $G^*{}_{ij} = G_{ada_{ij}}$
    end for
  end while
Output:  Final pseudo ground-truth boxes $G^*$
```



Input Images

pseudo ground-truth boxes adaptation (PGA)

c1

c2

Fully-Supervised Detector(FSD)

1. Use final pseudo ground-truth boxes
2. Based on RPN, train a **fully-supervised detector**

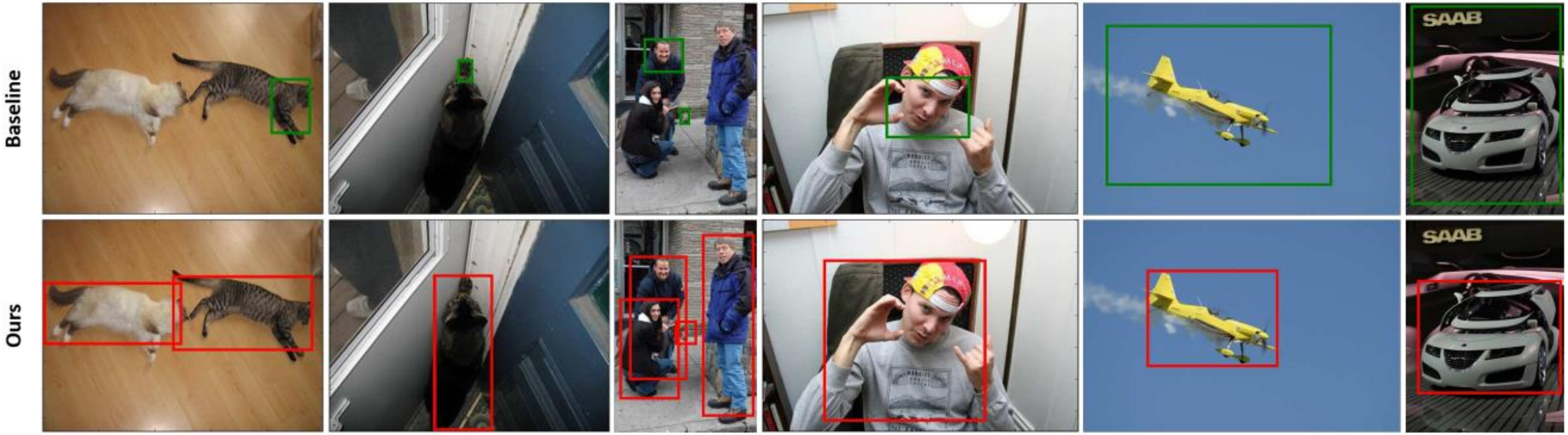
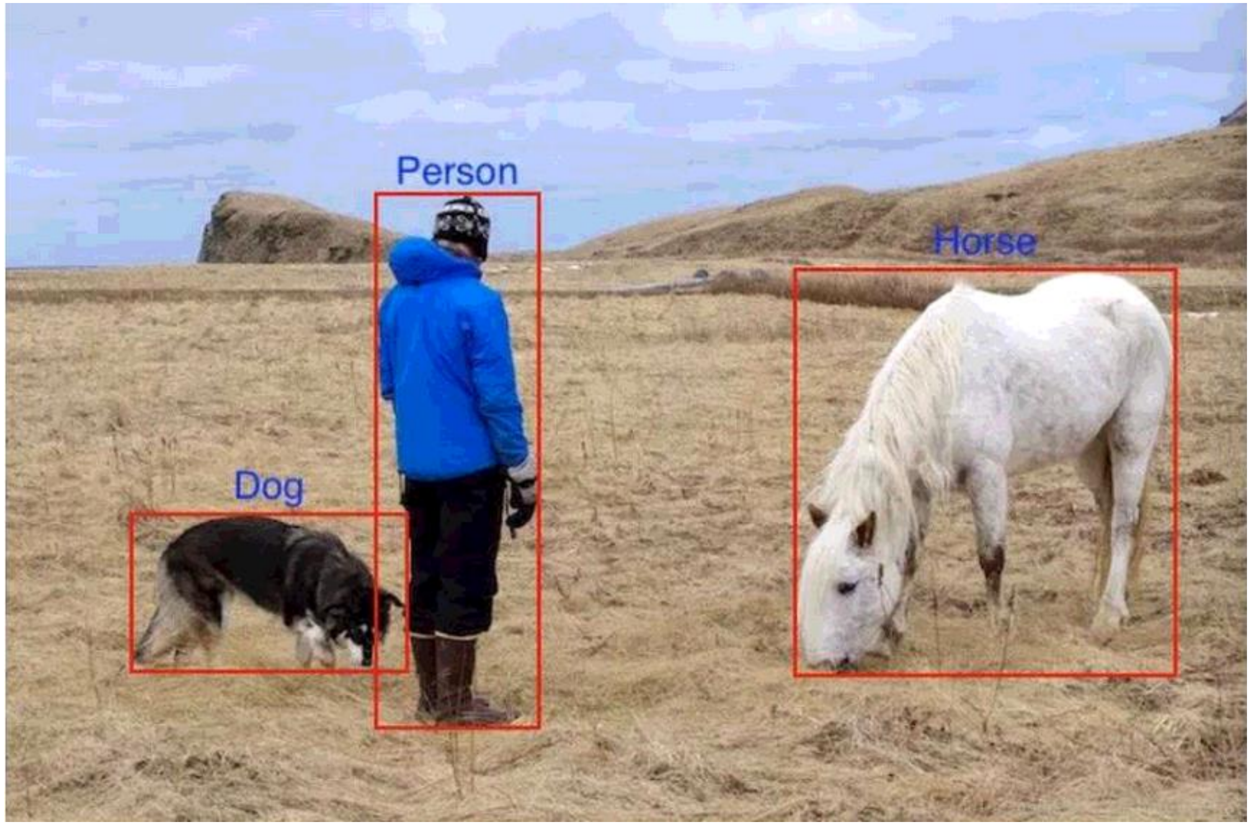


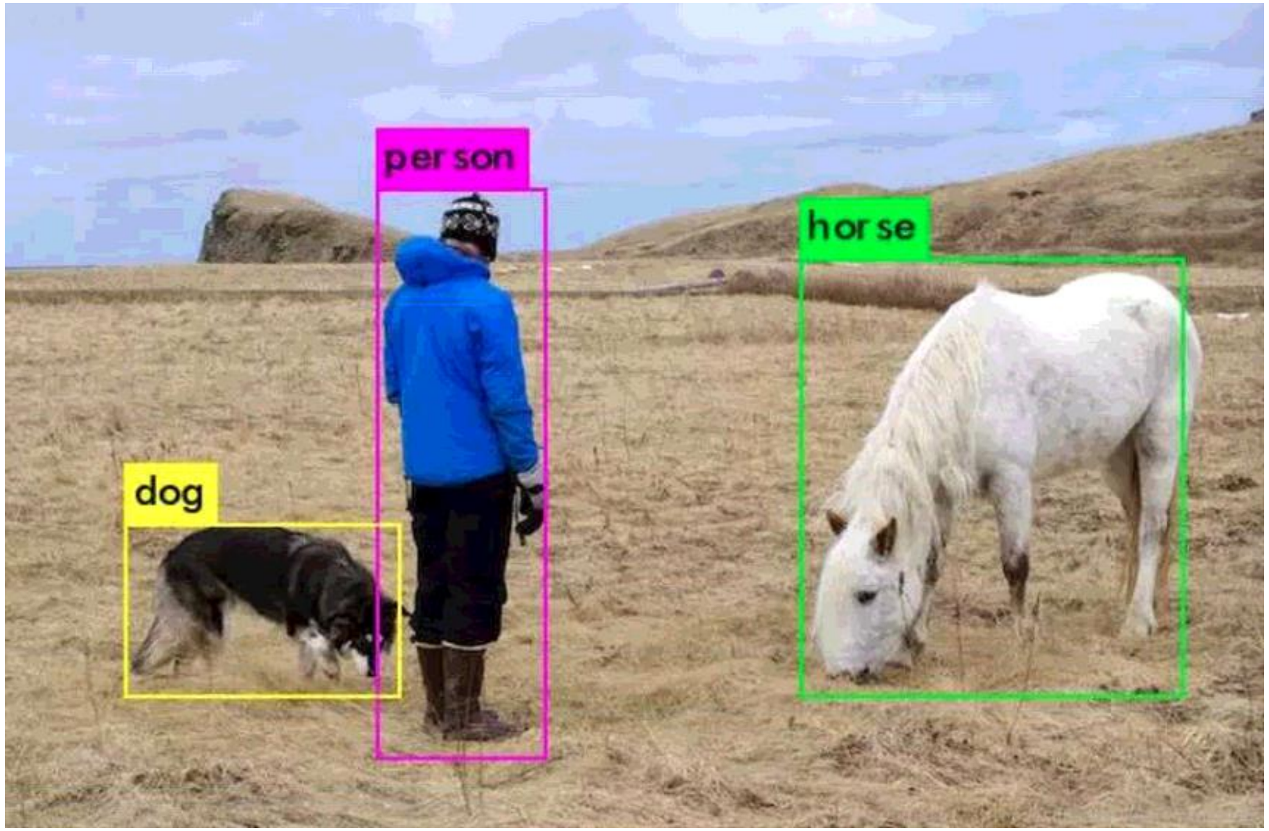
Figure 3. Some examples of pseudo ground-truth boxes generated by different weakly supervised detection methods. The top row shows the results of baseline [32] (*i.e.* selecting the top proposal with the highest predicted score as the pseudo ground-truth). The bottom row shows some pseudo ground-truth boxes mined by our method (*i.e.* PGE and PGA).

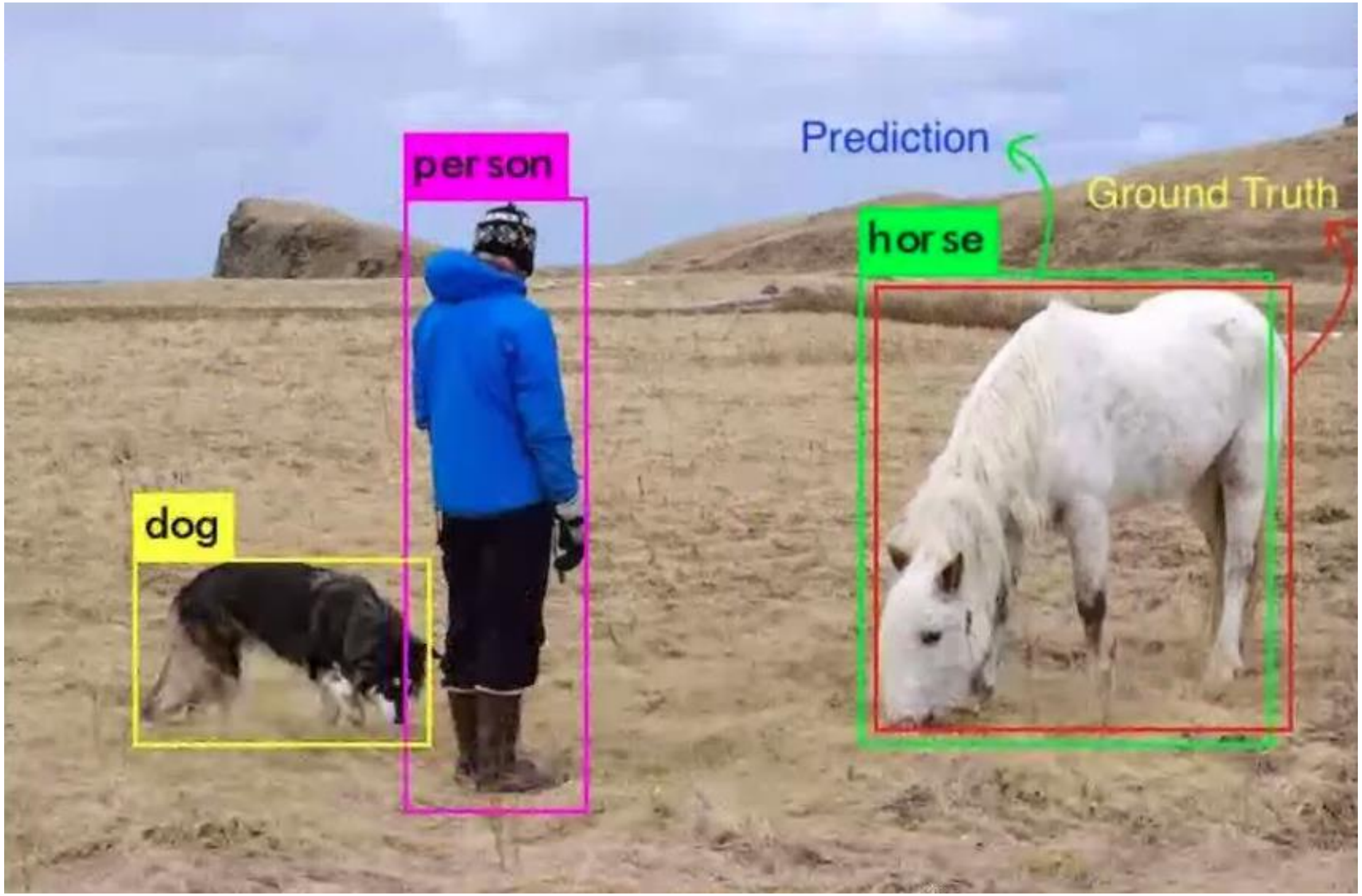
Experiments

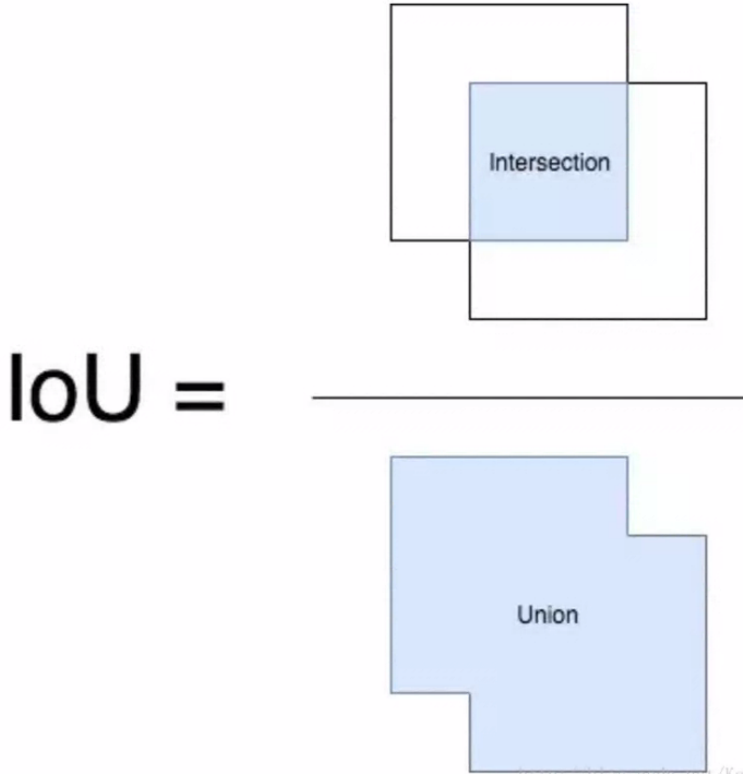
1. PASCAL VOC 2007 and 2012
2. 9,963 and 22,531 images from 20 object categories, respectively
3. Evaluation metrics : mean average precision (mAP)

mAP

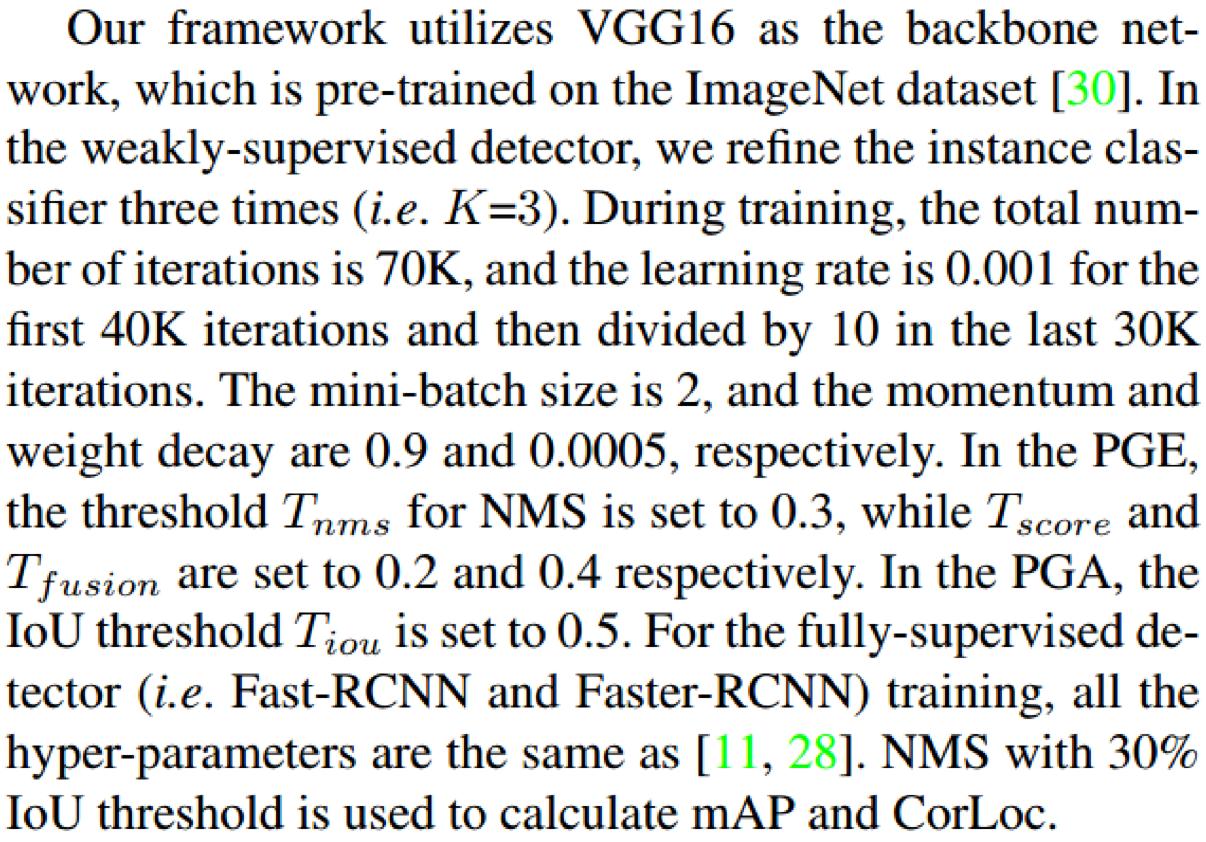




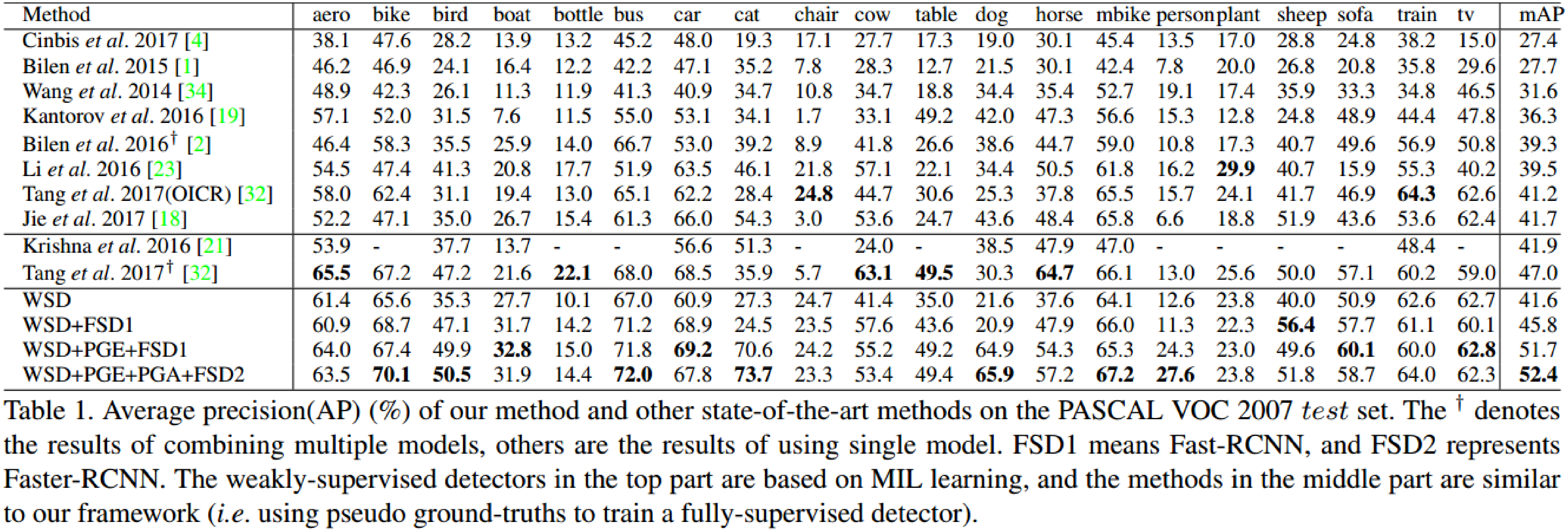

Prediction
Ground Truth
person
horse
dog


IoU =
Intersection
Union

Implementing details



Our framework utilizes VGG16 as the backbone network, which is pre-trained on the ImageNet dataset [30]. In the weakly-supervised detector, we refine the instance classifier three times (*i.e.* K=3). During training, the total number of iterations is 70K, and the learning rate is 0.001 for the first 40K iterations and then divided by 10 in the last 30K iterations. The mini-batch size is 2, and the momentum and weight decay are 0.9 and 0.0005, respectively. In the PGE, the threshold T_{nms} for NMS is set to 0.3, while T_{score} and T_{fusion} are set to 0.2 and 0.4 respectively. In the PGA, the IoU threshold T_{iou} is set to 0.5. For the fully-supervised detector (*i.e.* Fast-RCNN and Faster-RCNN) training, all the hyper-parameters are the same as [11, 28]. NMS with 30% IoU threshold is used to calculate mAP and CorLoc.



Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Cinbis *et al*. 2017 [4]	38.1	47.6	28.2	13.9	13.2	45.2	48.0	19.3	17.1	27.7	17.3	19.0	30.1	45.4	13.5	17.0	28.8	24.8	38.2	15.0	27.4
Bilen *et al*. 2015 [1]	46.2	46.9	24.1	16.4	12.2	42.2	47.1	35.2	7.8	28.3	12.7	21.5	30.1	42.4	7.8	20.0	26.8	20.8	35.8	29.6	27.7
Wang *et al*. 2014 [34]	48.9	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	18.8	34.4	35.4	52.7	19.1	17.4	35.9	33.3	34.8	46.5	31.6
Kantorov *et al*. 2016 [19]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	49.2	42.0	47.3	56.6	15.3	12.8	24.8	48.9	44.4	47.8	36.3
Bilen *et al*. 2016† [2]	46.4	58.3	35.5	25.9	14.0	66.7	53.0	39.2	8.9	41.8	26.6	38.6	44.7	59.0	10.8	17.3	40.7	49.6	56.9	50.8	39.3
Li *et al*. 2016 [23]	54.5	47.4	41.3	20.8	17.7	51.9	63.5	46.1	21.8	57.1	22.1	34.4	50.5	61.8	16.2	**29.9**	40.7	15.9	55.3	40.2	39.5
Tang *et al*. 2017(OICR) [32]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	**24.8**	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	**64.3**	62.6	41.2
Jie *et al*. 2017 [18]	52.2	47.1	35.0	26.7	15.4	61.3	66.0	54.3	3.0	53.6	24.7	43.6	48.4	65.8	6.6	18.8	51.9	43.6	53.6	62.4	41.7
Krishna *et al*. 2016 [21]	53.9	-	37.7	13.7	-	-	56.6	51.3	-	24.0	-	38.5	47.9	47.0	-	-	-	-	48.4	-	41.9
Tang *et al*. 2017† [32]	**65.5**	67.2	47.2	21.6	**22.1**	68.0	68.5	35.9	5.7	**63.1**	**49.5**	30.3	**64.7**	66.1	13.0	25.6	50.0	57.1	60.2	59.0	47.0
WSD	61.4	65.6	35.3	27.7	10.1	67.0	60.9	27.3	24.7	41.4	35.0	21.6	37.6	64.1	12.6	23.8	40.0	50.9	62.6	62.7	41.6
WSD+FSD1	60.9	68.7	47.1	31.7	14.2	71.2	68.9	24.5	23.5	57.6	43.6	20.9	47.9	66.0	11.3	22.3	**56.4**	57.7	61.1	60.1	45.8
WSD+PGE+FSD1	64.0	67.4	49.9	**32.8**	15.0	71.8	**69.2**	70.6	24.2	55.2	49.2	64.9	54.3	65.3	24.3	23.0	49.6	**60.1**	60.0	**62.8**	51.7
WSD+PGE+PGA+FSD2	63.5	**70.1**	**50.5**	31.9	14.4	**72.0**	67.8	**73.7**	23.3	53.4	49.4	**65.9**	57.2	**67.2**	**27.6**	23.8	51.8	58.7	64.0	62.3	**52.4**

Table 1. Average precision(AP) (%) of our method and other state-of-the-art methods on the PASCAL VOC 2007 *test* set. The † denotes the results of combining multiple models, others are the results of using single model. FSD1 means Fast-RCNN, and FSD2 represents Faster-RCNN. The weakly-supervised detectors in the top part are based on MIL learning, and the methods in the middle part are similar to our framework (*i.e*. using pseudo ground-truths to train a fully-supervised detector).

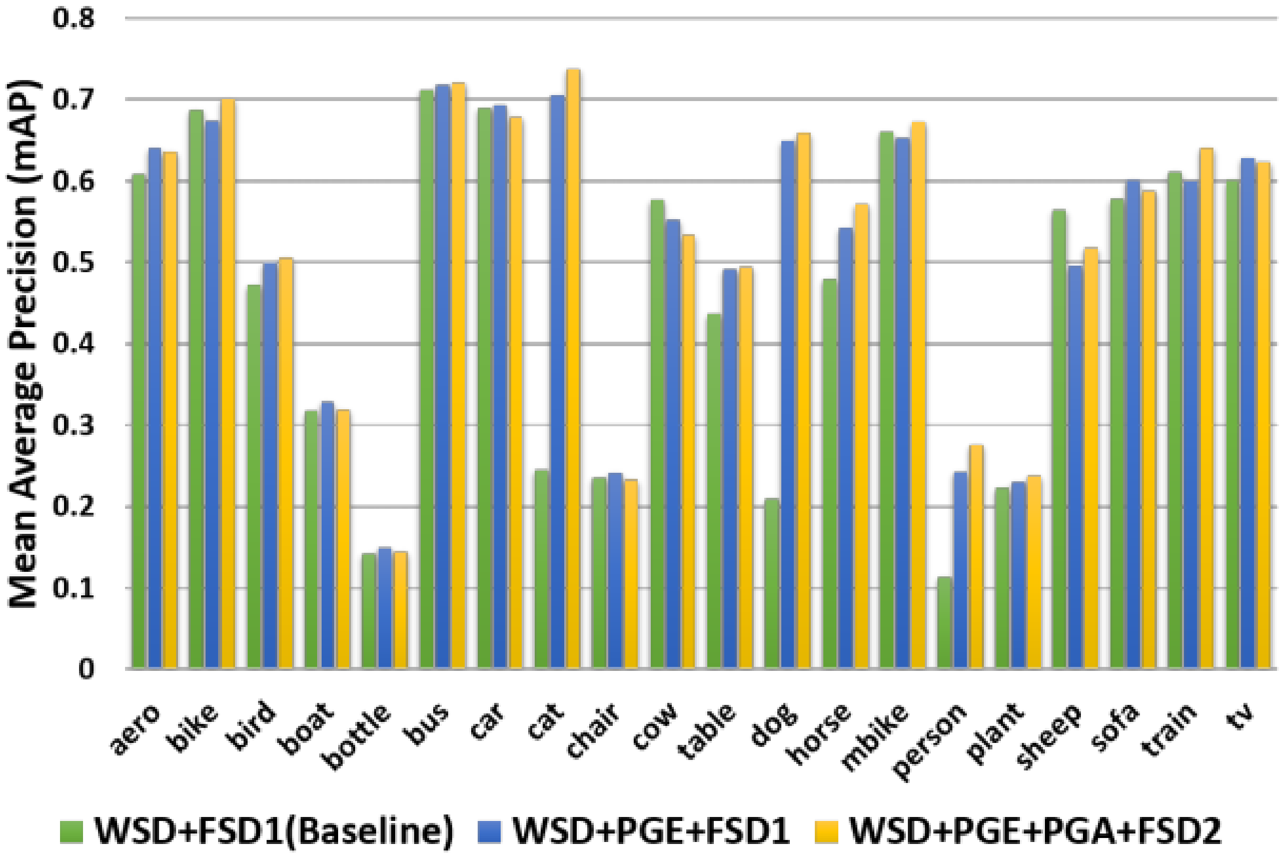
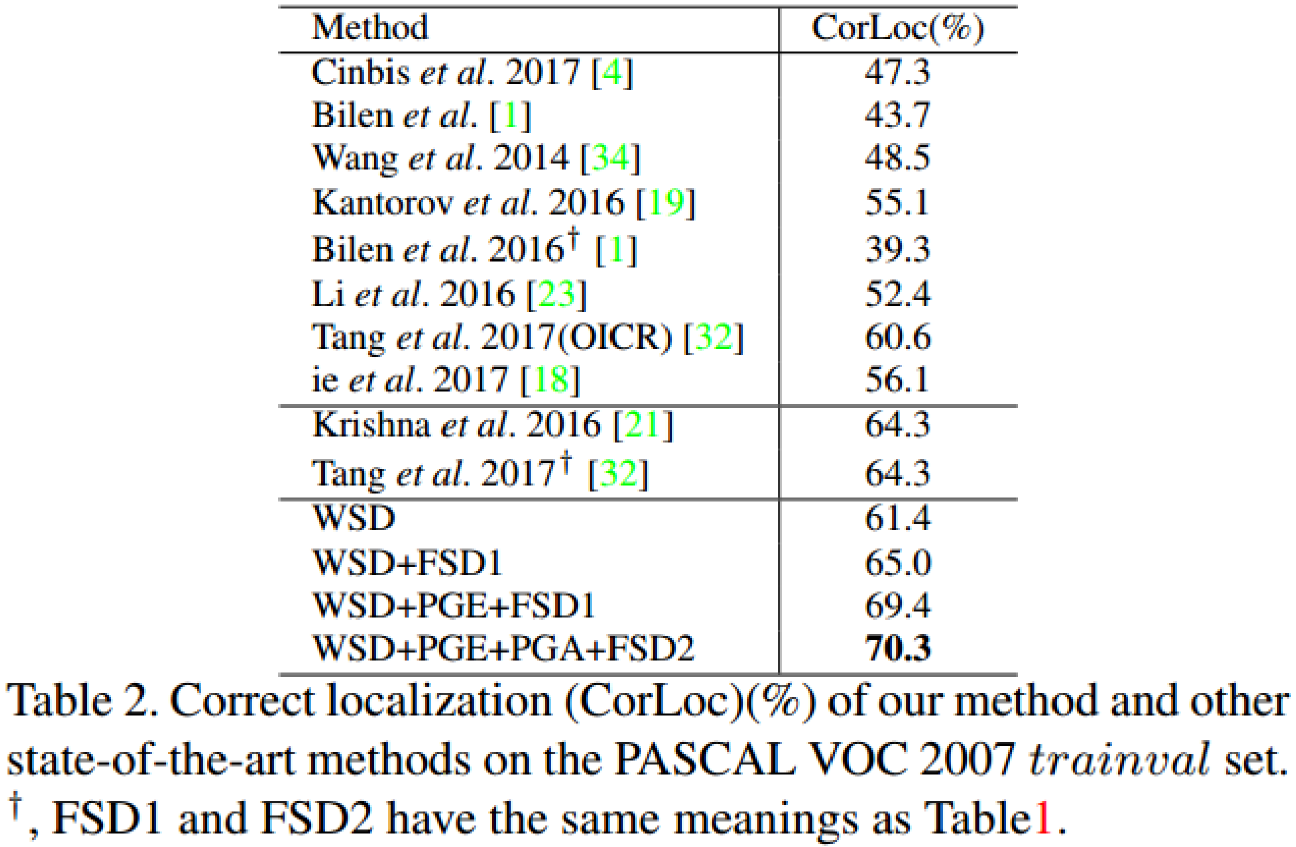


Figure 4. The mAP of each class in different ablation versions of our framework on VOC 2007 *test* set.



Method	CorLoc(%)
Cinbis *et al*. 2017 [4]	47.3
Bilen *et al*. [1]	43.7
Wang *et al*. 2014 [34]	48.5
Kantorov *et al*. 2016 [19]	55.1
Bilen *et al*. 2016† [1]	39.3
Li *et al*. 2016 [23]	52.4
Tang *et al*. 2017(OICR) [32]	60.6
ie *et al*. 2017 [18]	56.1
Krishna *et al*. 2016 [21]	64.3
Tang *et al*. 2017† [32]	64.3
WSD	61.4
WSD+FSD1	65.0
WSD+PGE+FSD1	69.4
WSD+PGE+PGA+FSD2	**70.3**

Table 2. Correct localization (CorLoc)(%) of our method and other state-of-the-art methods on the PASCAL VOC 2007 *trainval* set. †, FSD1 and FSD2 have the same meanings as Table1.

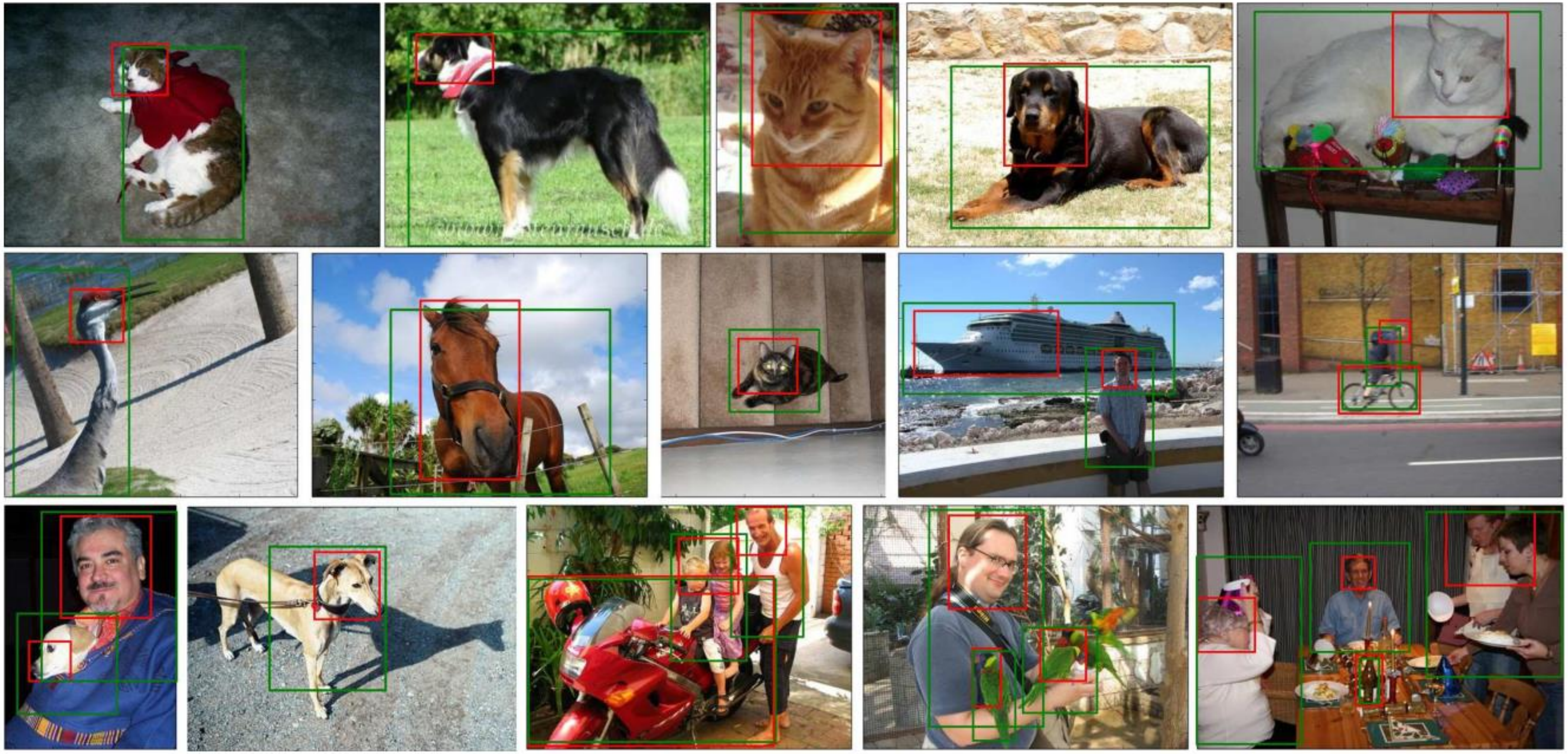
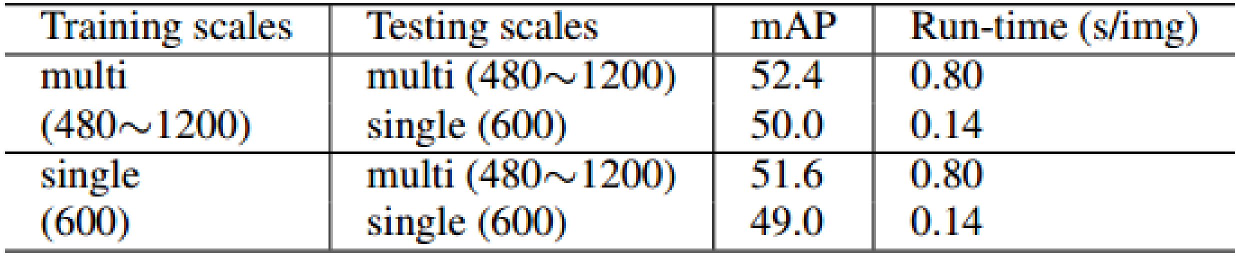
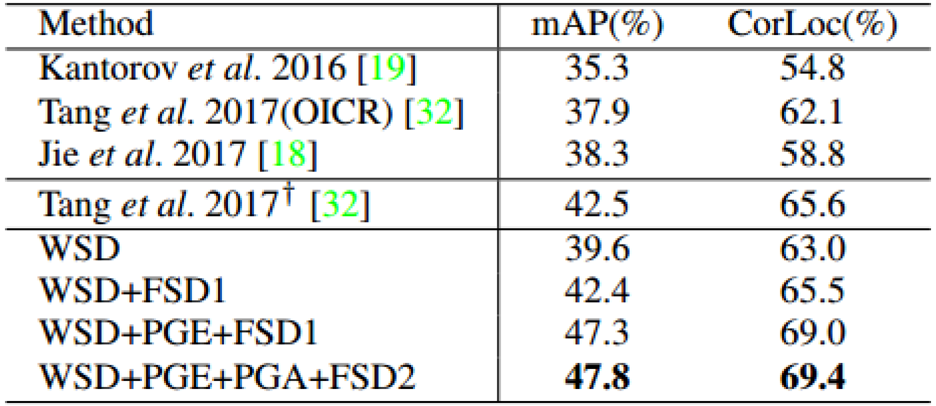


Figure 5. Qualitative detection results of our method (WSD+PGE+PGA+FSD2) and the baseline (WSD+FSD). Green bounding boxes indicate objects detected by our method, while red ones correspond to those detected by the baseline.



Training scales	Testing scales	mAP	Run-time (s/img)
multi (480∼1200)	multi (480∼1200)	52.4	0.80
	single (600)	50.0	0.14
single (600)	multi (480∼1200)	51.6	0.80
	single (600)	49.0	0.14

Table 3. The performance and the run-time of inference under different settings on PASCAL VOC 2007.



Method	mAP(%)	CorLoc(%)
Kantorov *et al*. 2016 [19]	35.3	54.8
Tang *et al*. 2017(OICR) [32]	37.9	62.1
Jie *et al*. 2017 [18]	38.3	58.8
Tang *et al*. 2017† [32]	42.5	65.6
WSD	39.6	63.0
WSD+FSD1	42.4	65.5
WSD+PGE+FSD1	47.3	69.0
WSD+PGE+PGA+FSD2	**47.8**	**69.4**

Table 4. Performance of our method and other state-of-the-art methods on the PASCAL VOC 2012. †, FSD1 and FSD2 have the same meanings as Table 1.

Conclusions

1. Combines the advantages of fully-supervised and weakly-supervised learning.
2. WSDNN and OICR to train a weakly-supervised detector (WSD) end-to-end.
3. And then by the virtue of PGE and PGA, finds high quality pseudo ground-truths from the WSD.
4. Finally, those pseudo ground-truths are fed into a fully-supervised detector to produce the final detection results.
5. Extensive experiments on PASCAL VOC 2007 and 2012 demonstrate the substantial improvements (5.4% and 5.3% in mAP respectively) of our method compared with previous state-of-the-art weakly-supervised detectors.