AttentionNet: Aggregating Weak Directions for Accurate Object Detection

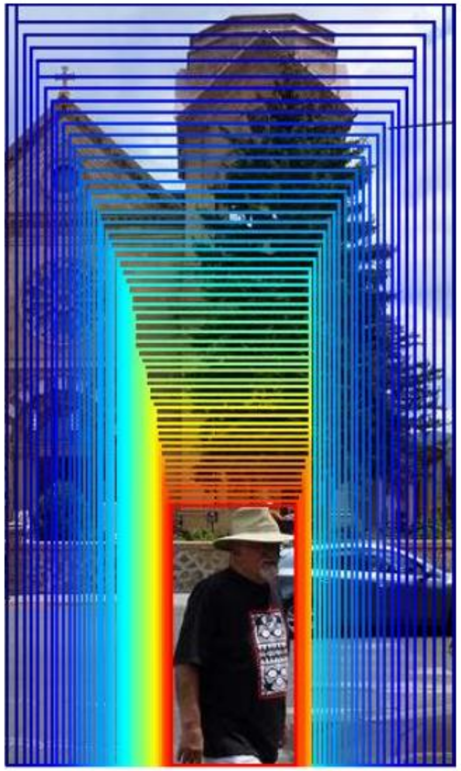
Donggeun Yoo, Sunggyun Park, Joon-Young Lee, Anthony S. Paek, ICCV 2015

报告人: YI Wu Kun 2018-12-31

Main idea

- Aggregation of many weak predictions.
- Objective: predict the top-left (TP) and bottom-right (BR) directions of the localization bounding box.
- Recursive crop of the input image.
- Multiple instance localization.






Stop signal.
Stop signal.
Stop signal.

letwork Architecture

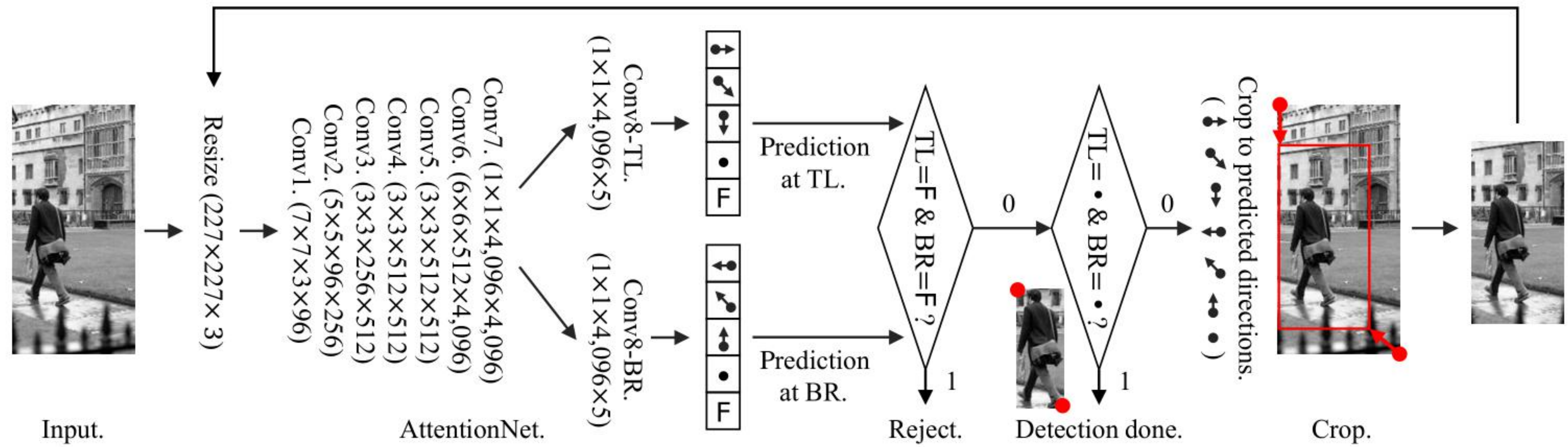
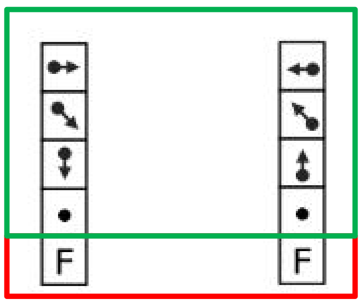


Figure 2. A pipeline of our detection framework. AttentionNet is composed of two final layers for top-left (TL) and bottom-right (BR) of the input image domain. Each of them outputs a direction (→ ↘ ↓ for TL, ← ↖ ↑ for BR) where each corner of the image should go to for the next step, or a "stop" sign (•), or "non-human" sign (F). When AttentionNet outputs "non-human" in both layers, the image is rejected. The image is cropped according to the weak directions and fed to AttentionNet again, until it meets "stop" in both layers.

raining

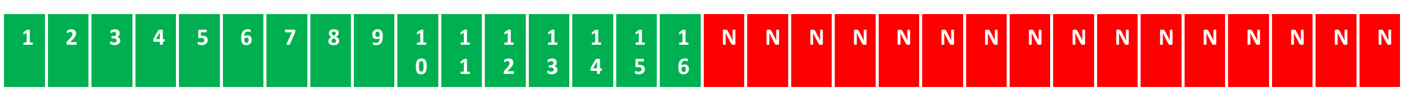
When we compose a batch to train the CNN, we select positive and negative regions in an equal portion. In a batch, each of the 16(=4 × 4) cases for positive regions occupies a portion of 1/(2 × 16), and the negative regions occupy the re-maining portion of 1/2. The loss for training AttentionNet is an average of the two soft-max losses computed indepen-dently in TL and BR.



4 * 4 = 16 cases of positive regions

1 * 1 = 1 cases of negative regions

Batch



OSS Function

2 softmax loss for TL & BR

$$S_{TLj} = \frac{e^{a_j}}{\sum_{k=1}^{5} e^{a_k}}$$

$$LOSS_{TL} = -\sum_{k=1}^{5} y_i \log(S_{TLj})$$

$$S_{BRj} = \frac{e^{a_j}}{\sum_{k=1}^{5} e^{a_k}}$$

$$LOSS_{BR} = -\sum_{k=1}^{5} y_i \log(S_{BRj})$$

$$LOSS = LOSS_{TL} + LOSS_{BR}$$

ugmentation(region)

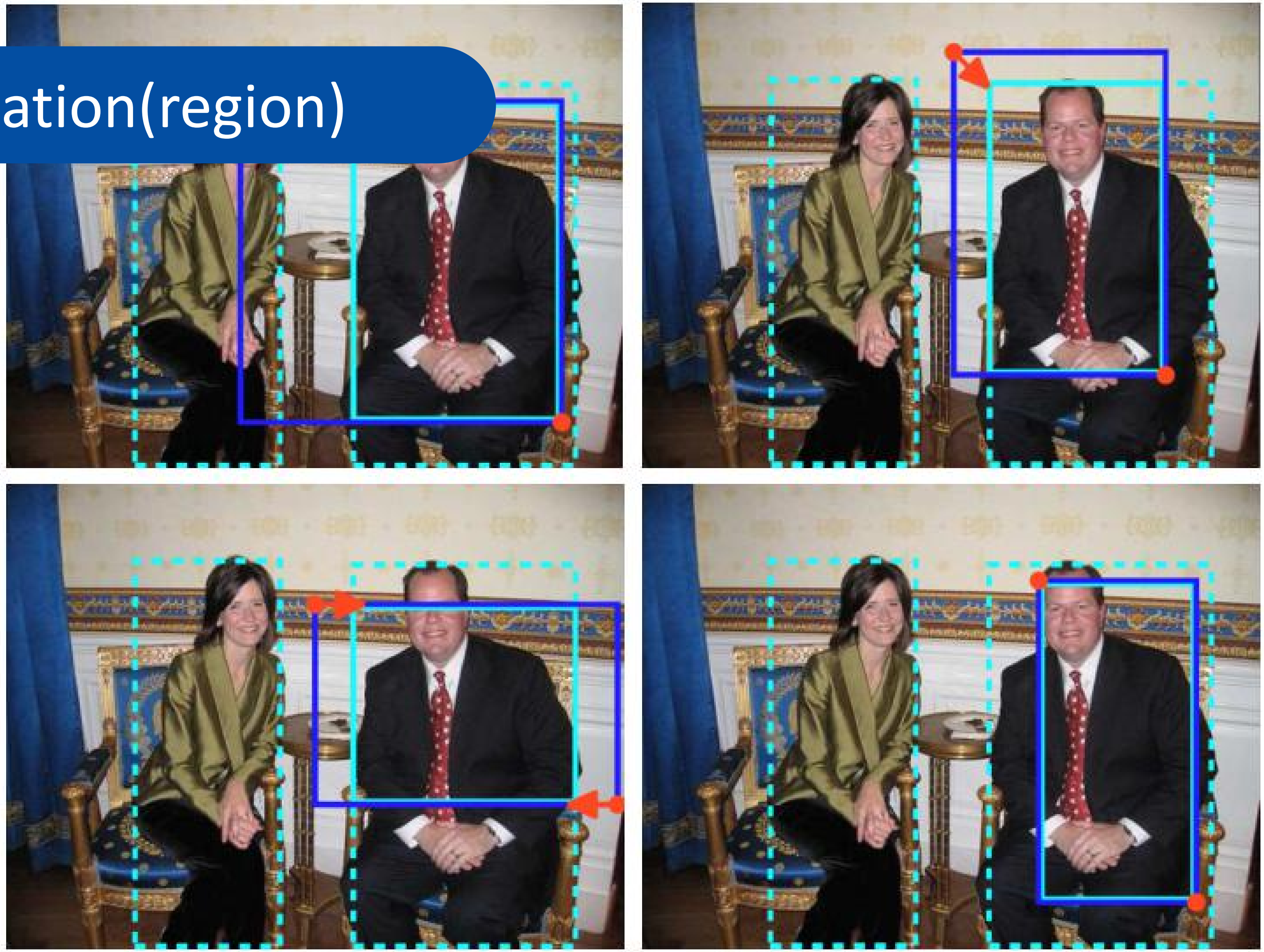


Figure 3. Real examples of crop-augmentation for training AttentionNet. The target instance is the right man. Dashed cyan bounding boxes are ground-truths, and the blue bounding boxes are the augmented regions. Red arrows/dots denote their ground truths.

ugmentation(region)

We randomly generate positive regions which satisfy the following three rules.

1. A positive region must include at least 50% of the area of a target instance.

2. A positive region can include multiple instances (as the top-left example in Fig. 3), but the target instance must occupy the biggest area. Within a cropped region, the area of the target instance must be at least 1.5-times larger than that of the other instances. The second rule is important for complex instance layouts in the multiple instance scenario (to be introduced in Sec. 4). Without this rule in the scenario, a final bounding box is prone to fit multiple instances at once. In order to make AttentionNet always narrow the bounding box down to the largest instances among multiple instances, we must follow the second rule in generating positive regions.

3. Regions are cropped in varying aspect ratios as well as varying scales.

Detection process

During the test stage, starting from an initial test over the entire image boundary to a final decision of “stop” or “no instance”, the number of possible decision pairs is 17 (=4 × 4+1) such as {→,&,↓,•} TL ×{←,-,↑,•} BR for positive regions and {F TL ,F BR } for negative regions.

L = 30px
Max iterative feed-forward = 50

$$s^b = s^b_{\mathrm{TL}} + s^b_{\mathrm{BR}}, \quad \text{s.t.}$$

$$s^b_{\mathrm{TL}} = y^{\bullet}_{\mathrm{TL}} - (y^{\rightarrow}_{\mathrm{TL}} + y^{\searrow}_{\mathrm{TL}} + y^{\downarrow}_{\mathrm{TL}} + y^{\mathrm{F}}_{\mathrm{TL}}),$$

$$s^b_{\mathrm{BR}} = y^{\bullet}_{\mathrm{BR}} - (y^{\leftarrow}_{\mathrm{BR}} + y^{\nwarrow}_{\mathrm{BR}} + y^{\uparrow}_{\mathrm{BR}} + y^{\mathrm{F}}_{\mathrm{BR}}).$$

Object Proposal method vs AttentionNet

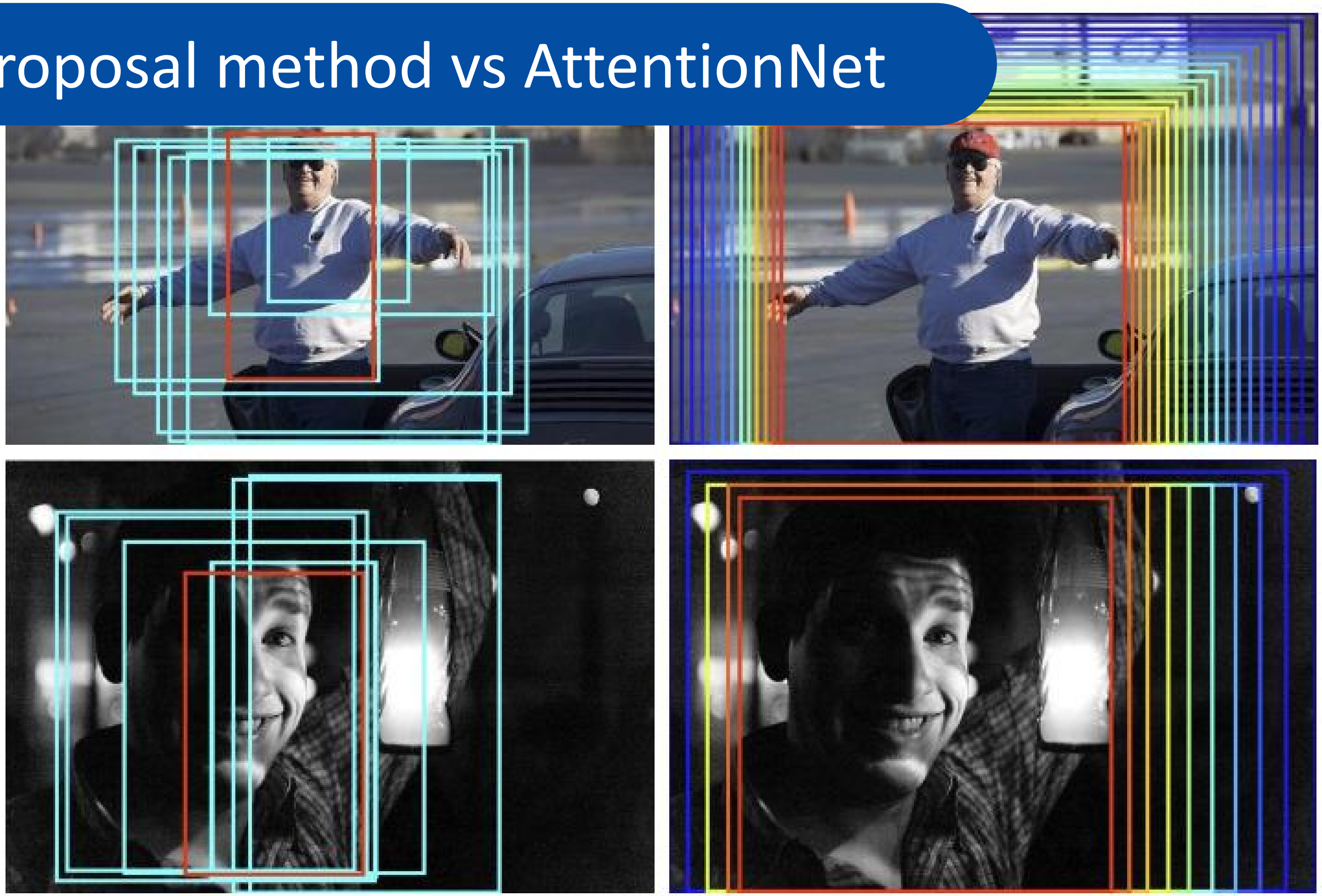


Figure 5. Real detection examples of the object proposal based method (left) and AttentionNet (right). In the left column, a red bounding box is the top-1 detected region among top-10 object proposals (cyan) with the maximum SVM score.

Multi Object detection

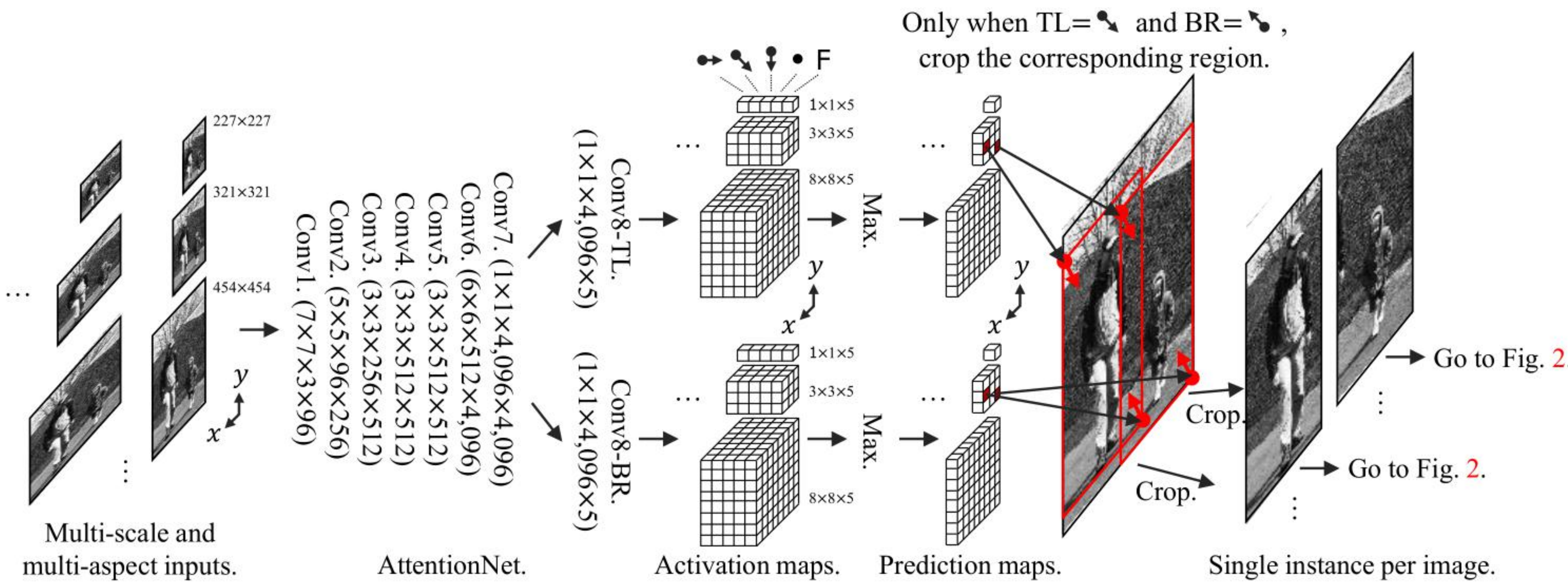
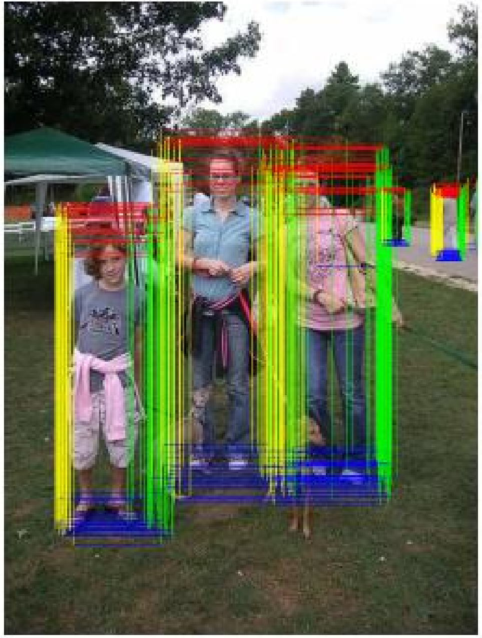
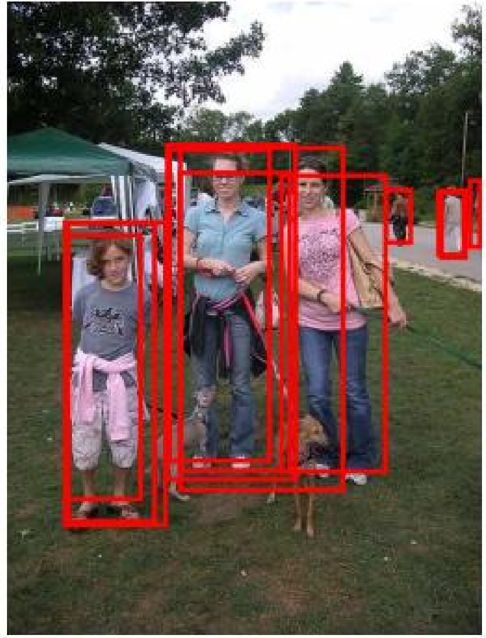


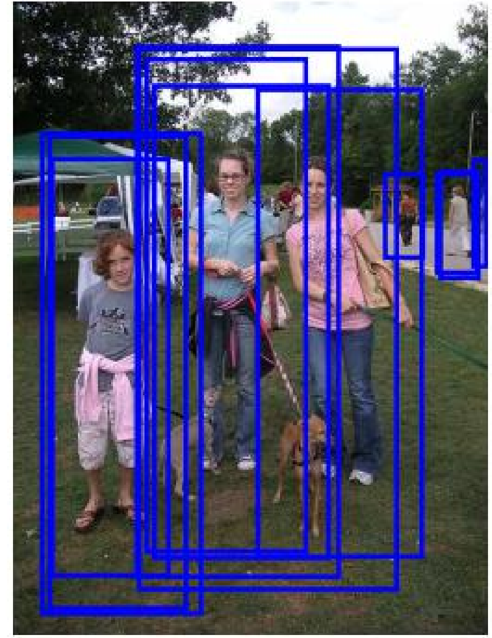
Figure 4. Extracting single-instance regions, where a single instance is included only. Multiple inputs with multiple scales/aspects are fed to AttentionNet, and prediction maps are produced. Only image regions satisfying $\{\searrow_{TL}, \nwarrow_{BR}\}$ are regarded as the single-instance regions. These regions are fed to AttentionNet again for final detection. Note, the CNN here and that in Fig. 2 are the same one, not separated.



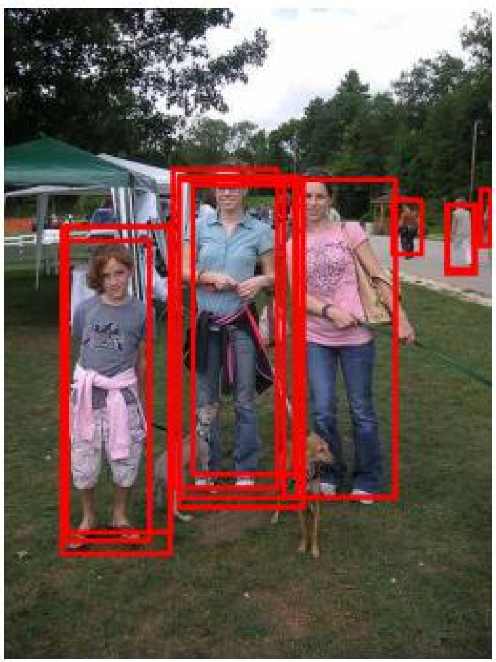
(a) Initial detections.



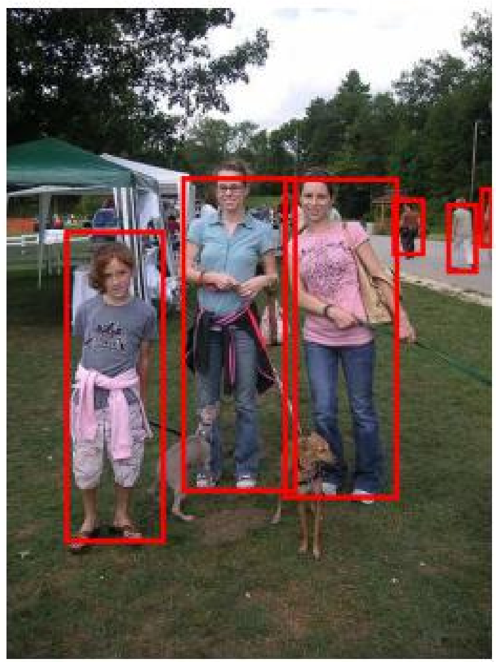
(b) Initial merge.



(c) Re-initialize. (×2.5)



(d) Re-detections.



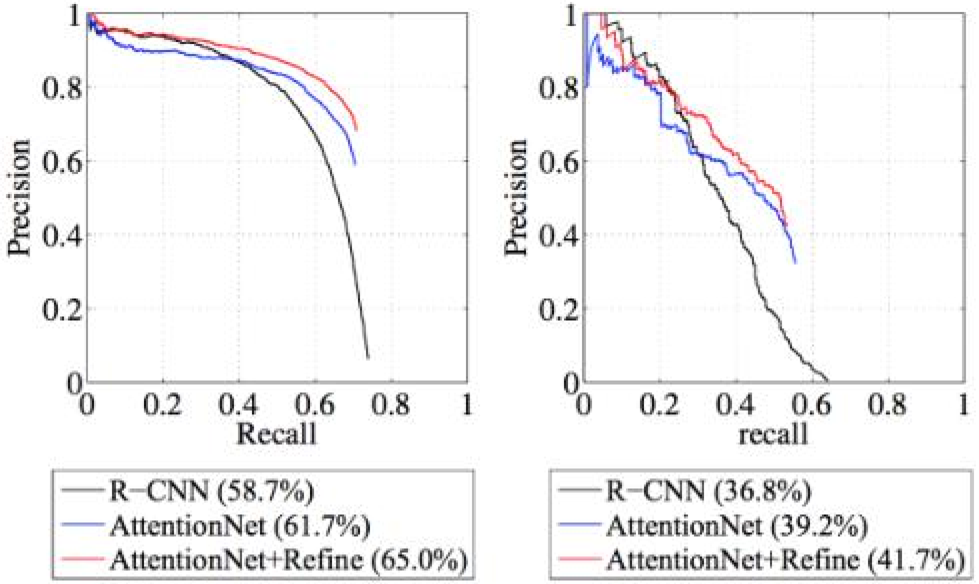
(e) Final merge.

Real examples of our detection procedure, including initial results (a~ b) and refinement (c~ e). Initially detected candidates come from Fig. 4 followed by Fig. 2 are merged by an intersection over union (IoU) of 0.8. We extend each merged box to 2.5-times larger
size, and feed them to Fig. 2 again. Finally we merge the second results by an IoU of 0.5.

Evaluation
(PASCAL VOC 2007/2012, AP reported)

Method	Extra data	VOC'07	VOC'12
AttentionNet	ImNet	61.7	62.8
AttentionNet + Refine	ImNet	**65.0**	**65.6**
AttentionNet + R-CNN	ImNet	66.4	69.0
AttentionNet + Refine + R-CNN	ImNet	**69.8**	**72.0**
Person R-CNN + BBReg	ImNet	59.7	N/A
Person R-CNN + BBReg×2	ImNet	**59.8**	N/A
Person R-CNN + BBReg×3	ImNet	59.7	N/A
Felzenszwalb *et al.*'10 [11]	None.	41.9	N/A
Bourdev *et al.*'10 [2]	H3D	46.9	N/A
Szegedy *et al.*'13 [24]	VOC'12	26.2	N/A
Erhan *et al.*'14 [9]	None.	37.5	N/A
Gkioxari *et al.*'14 [13]	VOC'12	45.6	N/A
Bourdev *et al.*'14 [3]	ImNet + H3D	59.3	58.7
He *et al.*'14 [14]	ImNet	57.6	N/A
Girshick *et al.*'14 [12]	ImNet	58.7	57.8
Girshick *et al.*'14 [12]	ImNet	**64.2***	N/A
Shen and Xue '14 [20]	ImNet	59.1	**60.2**

*Very deep model of 16 convolution layers [21] is used.



(a) Person class (b) Bottle class

ontributions

1. We suggest a novel detection method, which estimates an exact bounding box by aggregating weak predictions from attentionNet.

2. Our method does not include any separated models such as the object proposal, object classifiers and post bounding box regressor. AttentionNet does all these.

3. We achieve the state-of-the-art performance on single-class object detection tasks.

imitations

1) Single Class only

2) low recall. Multi object detection use $\{\searrow_{TL}, \nwarrow_{BR}\}$ to generate the region proposal and continue to input to AttentionNet to detect again.