Structure Inference Net: Object Detection Using Scene-Level Context and Instance-Level Relationships

Yong Liu[1,2], Ruiping Wang[1,2,3], Shiguang Shan[1,2,3], Xilin Chen[1,2,3]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Cooperative Medianet Innovation Center, China
yong.liu@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

This document provides more quantitive results of localization, occlusion and area size, more qualitative results of relationships and more detection samples in the following sections, where the baseline method is Faster R-CNN.

1. Localization

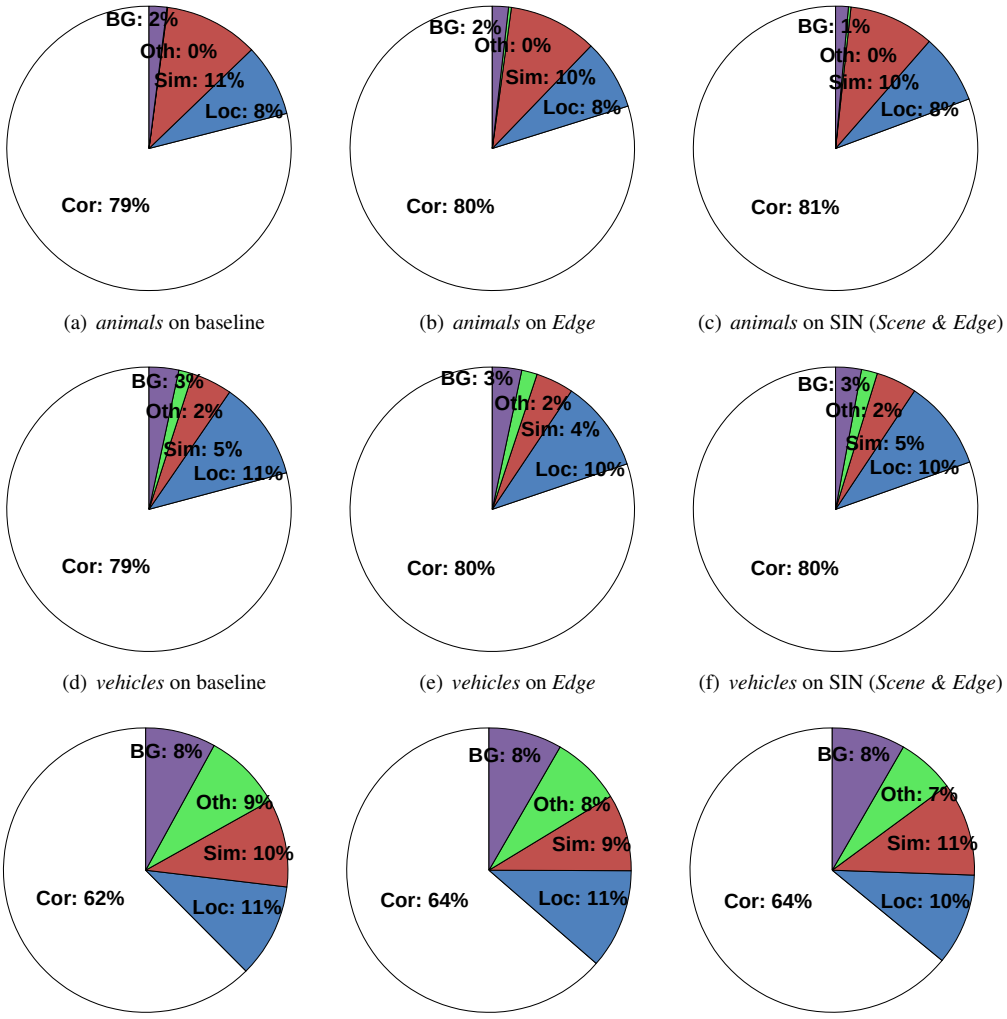
This section gives additional results of Sec. 5.2 (**Localization** part) in the main paper. The 20 categories in PASCAL VOC are divided into three super-categories including *animals, vehicles* and *furnitures*. Fig. 1 takes these super-categories to show the frequency and impact on the performance of each type of false positive. One can see that *Edge* has fewer localization errors on *vehicles* compared with the baseline, and SIN performs best.

2. Occlusion & Area Size

This section details the occlusion and area size analysis of SIN. We inspect the performance variations for each characteristic on seven categories selected by [1] (*i.e.* Ref. [20] in the main paper) on PASCAL VOC 2007. Here, results on three typical categories including *boat, chair* and *dining table* are presented in Fig. 2. We can learn that SIN performs better with occluded, truncated and small objects compared with the baseline.

3. Relationships Visualization

This section gives more qualitative results of relationships produced by SIN to check whether the relative object-object relationship has really been learned on COCO in Fig. 3.



(a) *animals* on baseline (b) *animals* on *Edge* (c) *animals* on SIN (*Scene & Edge*)

(d) *vehicles* on baseline (e) *vehicles* on *Edge* (f) *vehicles* on SIN (*Scene & Edge*)

(g) *furnitures* on baseline (h) *furnitures* on *Edge* (i) *furnitures* on SIN (*Scene & Edge*)

Figure 1. **Analysis of Top-Ranked False Positives.** Pie charts: fraction of detections that are correct (Cor) or false positive due to poor localization (Loc), confusion with similar objects (Sim), confusion with other VOC objects (Oth), or confusion with background or unlabeled objects (BG). In every pair of results, the **left** is based on baseline, the **middle** is based on *Edge* and the **right** is based on SIN. Loc errors of our SIN method are fewer than the baseline.

4. Qualitative Results on VOC

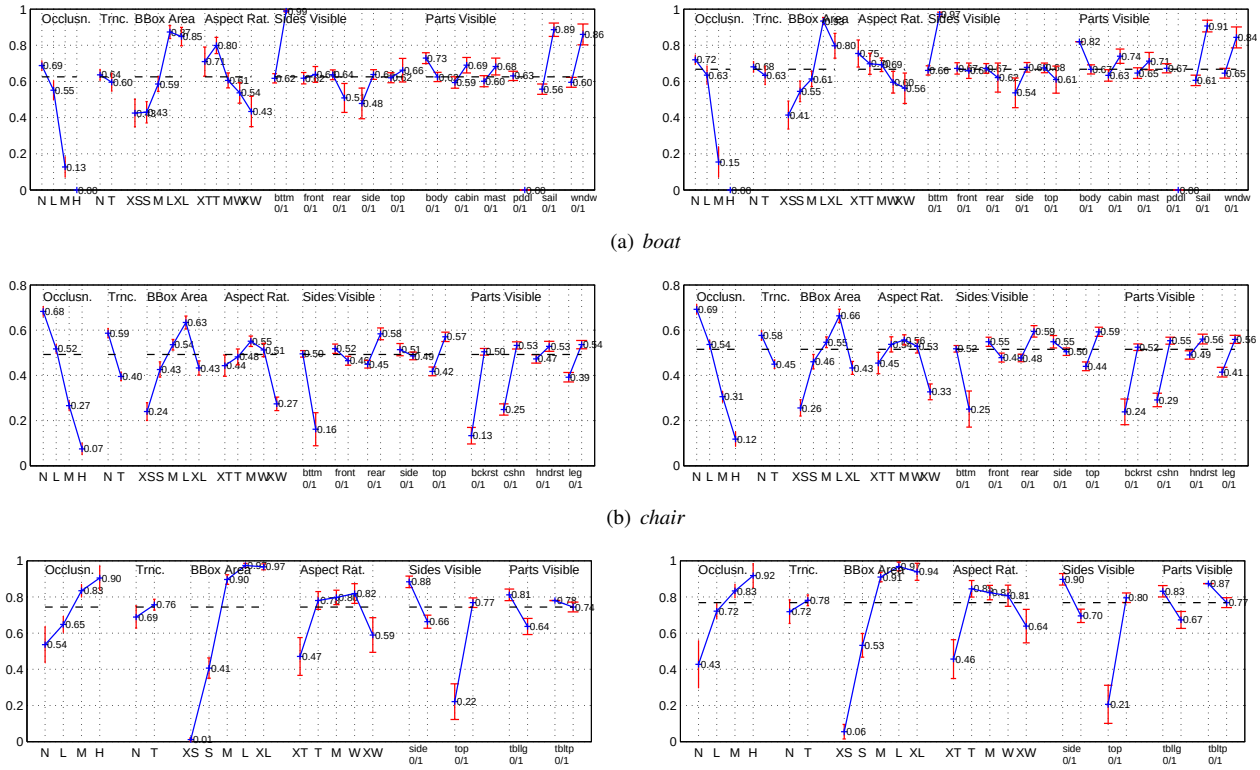
In this part, we present more qualitative results of SIN versus the baseline on VOC in Fig. 4.

5. Qualitative Results on COCO

In this part, we present more qualitative results of SIN versus the baseline on COCO in Fig. 5.

References

[1] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *ECCV*, 2012. 1



(a) *boat*

(b) *chair*

(c) *dining table*

Figure 2. **Per-Category Analysis of Characteristics.** AP_N (+) with standard error bars (red). Black dashed lines indicate overall AP_N. Key: **Occlusion**: N=none; L=low; M=medium; H=high. **Truncation**: N=not truncated; T=truncated. **Box Area**: XS=extra-small; S=small; M=medium; L=large; XL=extra-large. **Aspect Ratio**: XT=extra-tall/narrow; T=tall; M=medium; W=wide; XW=extra-wide. **Part Visibility / Viewpoint**: 1=part/side is visible; 0=part/side is not visible. In every pair of results, the **left** is based on baseline, and the **right** is detection result of SIN. We can learn that SIN performs better with occluded, truncated and small objects compared with the baseline.

Figure 3. **More Relative Objects Visualization on COCO.** It is learned that those objects connected by red dashed line are most relative.

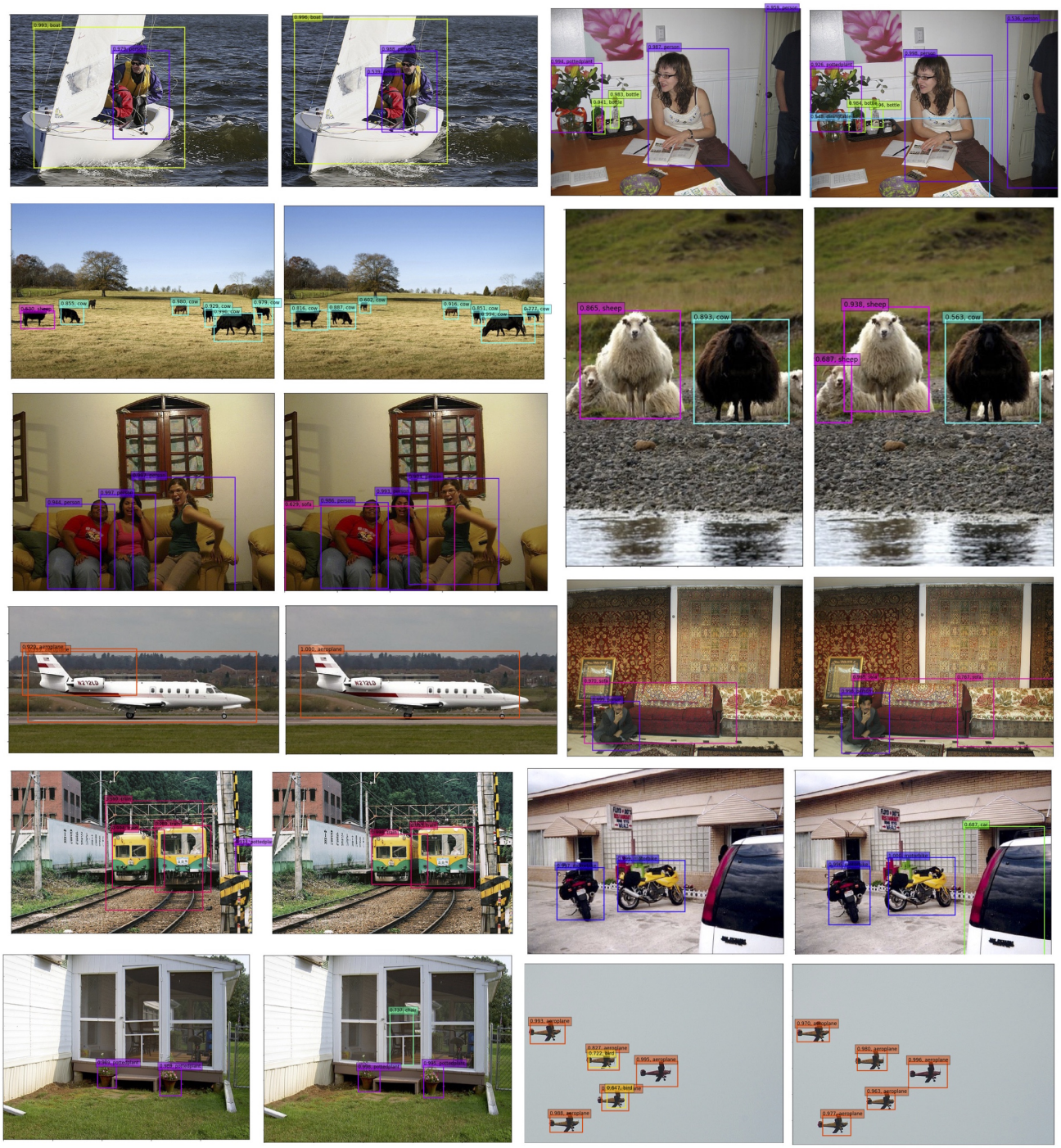


Figure 4. **More Qualitative results of Baseline vs. SIN on VOC.** In every pair of detection results, the **left** is based on baseline, and the **right** is detection result of SIN. We can see that SIN always performs better.

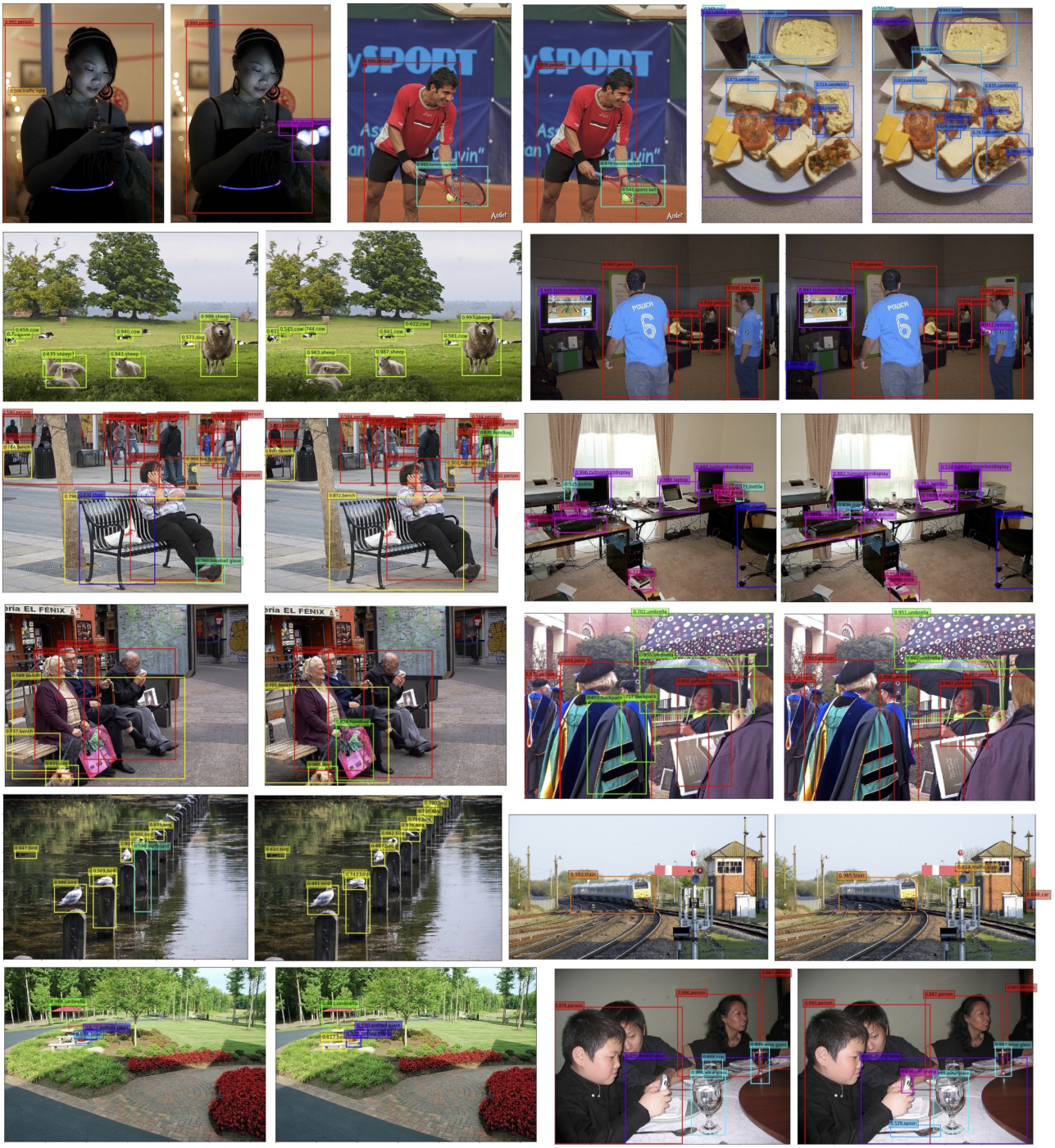


Figure 5. **More Qualitative results of Baseline vs. SIN on COCO.** In every pair of detection results, the **left** is based on baseline, and the **right** is detection result of SIN. We can see that SIN always performs better.