Dynamic Refinement Network for Oriented and Densely Packed Object Detection – Supplementary Materials

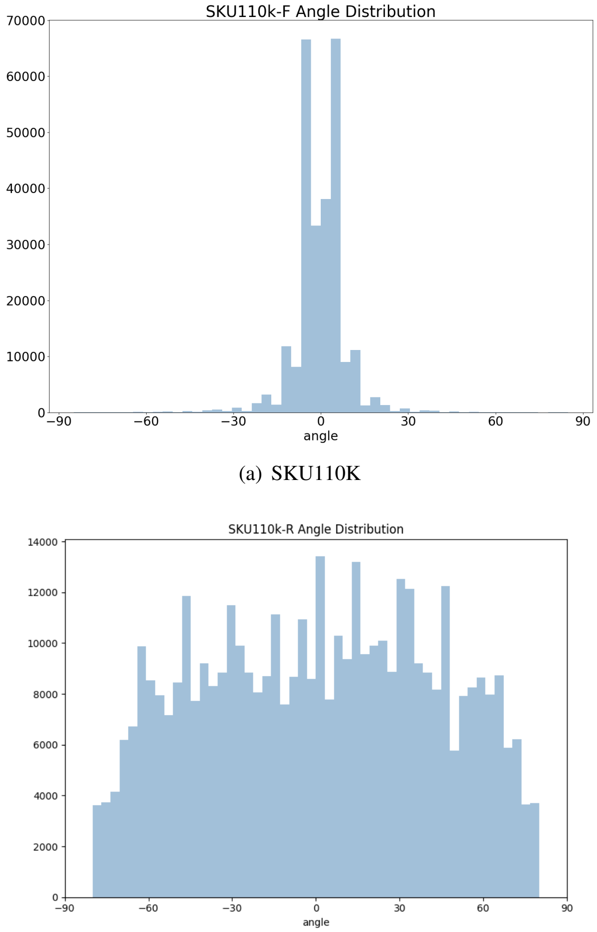
Xingjia Pan[1,2] Yuqiang Ren[3] Kekai Sheng[3] Weiming Dong[1,2,4]*
Haolei Yuan[3] Xiaowei Guo[3] Chongyang Ma[5] Changsheng Xu[1,2,4]
[1]NLPR, Institute of Automation, CAS [2]School of Artificial Intelligence, UCAS
[3]Youtu Lab, Tencent [4]CASIA-LLVision Joint Lab [5]Y-Tech, Kuaishou Technology
{panxingjia2015, weiming.dong, changsheng.xu}@ia.ac.cn, chongyangma@kuaishou.com
{condiren, saulsheng, harryyuan, scorpioguo}@tencent.com

1. SKU110K-R

Popular object detection sets include ILSVRC [1], PASCAL VOC [3] detection challenges, MS COCO [6] and the very recent Open Images v4 [5]. None of these provides scenes with packed items. A number of recent benchmarks emphasize crowded scenes, but are designed for counting, rather than detection [4]. SKU110K is a new dataset proposed by Goldman [4] which focus on retail environments. The images were collected from thousands of supermarket stores around the world. It is characterized by densely packed and slightly oriented items. Each instance is annotated with a horizontal bounding box.

On the basis of SKU110K, we propose an extensive variant, namely SKU110K-R, of which each instance is annotated by an oriented bounding box. In the original SKU110K, the orientation angle ranges mainly in [-15°, 15°]. To enrich the orientation, we further do some rotation augmentation from 6 angles (-45°, -30°, -15°, 15°, 30° ,45°). Fig. 1 shows the statistics of orientation distribution of instances in SKU110k and SKU110K-R. To be compatible with the setting of CenterNet, we use a tuple(cx,cy,w,h,θ) to depict a oriented bounding box. cx,cy are the coordinates of the center point. w,h are the width and height of the object and θ is the orientation angle. Note that we start with y-axis, positive in clockwise direction and negative in counterclockwise direction. All the angles ranges from -90° to 90°.

We have a professional and skilled labeling team responsible for annotating oriented bounding boxe for each instance. To maintain consistent with original dataset, there is no bounding box added or deleted during the annotation process, so each oriented bounding box of SKU110K-R has a corresponding horizontal bounding box of SKU110K. Fig. 2 shows some examples with annotated oriented and



(a) SKU110K

(b) SKU110K-R

Figure 1. Statistics of orientation distribution of SKU110K and SKU110K-R. Y axis represents the number of instance and X axis represents the orientation angle.

*Corresponding author

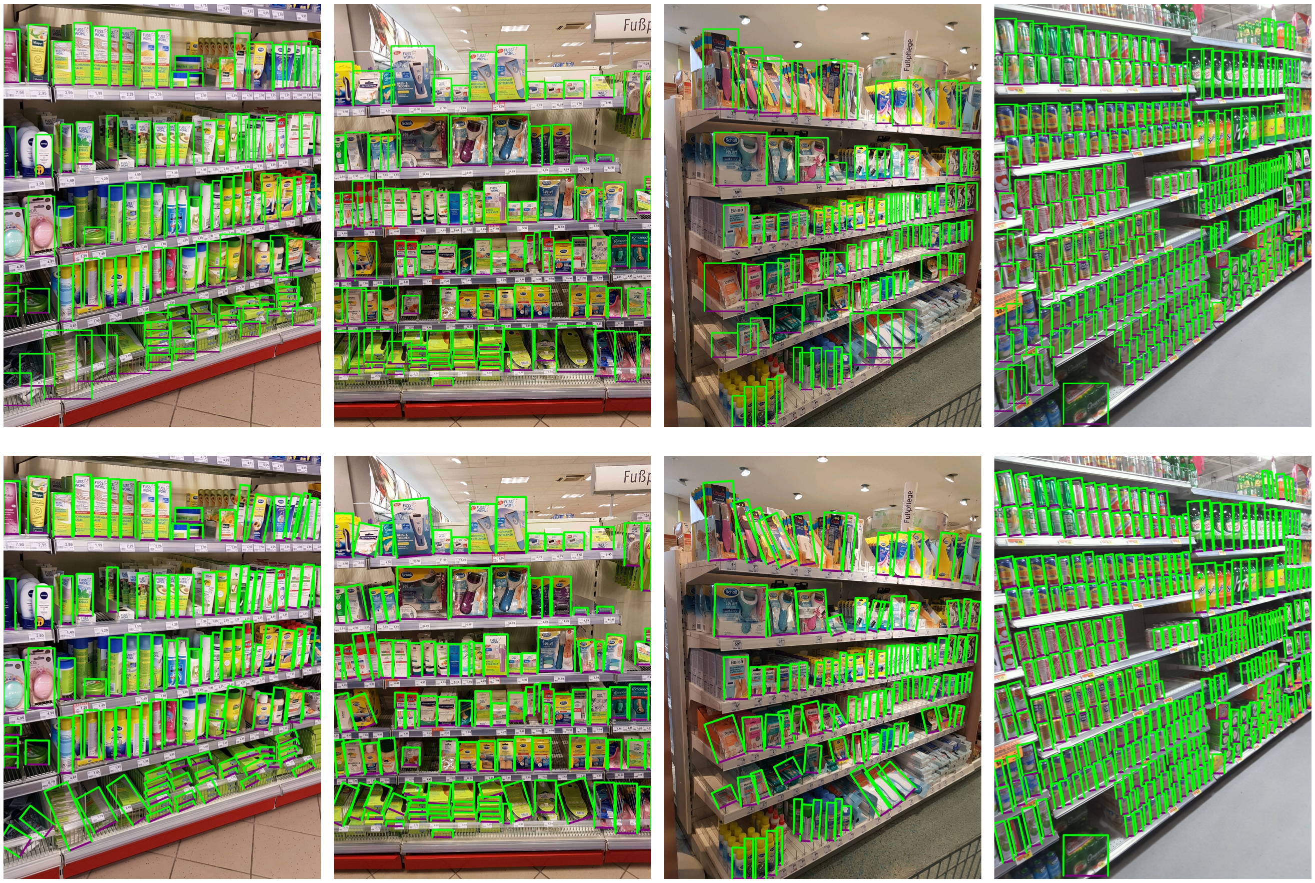


Figure 2. Some sample images from SKU110K. The images in **top** row are annotated with horizontal bounding boxes while the images in **bottom** row are with oriented bounding boxes.

horizontal bounding boxes.

2. Implementation Details

CenterNet†. The framework of CenterNet† is shown in Fig. 3. The main difference from original CenterNet is that we insert a deformable center pooling layer [2] between backbone and heads. Center pooling layer is proposed in [2], and we replace the standard convolution layer with deformable convolution layer.

CenterNet-4point†. The only difference from CenterNet† is that we regress the four corners of oriented bounding boxes directly instead of regressing the width and height of objects.

3. Visualization Results

We visualize two examples of attention map in FSM. The attention map guides the model to adjust the kernel. Fig. 4 illustrates two attention maps. Note that we only plot the attention map at the regions where the predict scores on heatmaps are larger than 0.05. Wathet blue, orange red, green and blue are responsible for kernel with 3×1, 5×1, 1×3 and 3×3, respectively. In (a), the majority of objects are “Large vehicle“ and with slender shape. Correspondingly, the kernel shapes in (a) for each object are 3×1(wathet blue) and 5×1(orange red). In (b), most objects belong to category “swimming pool“ and of which the shapes are likely with square or flat. Correspondingly, the kernels are with 1×3(green) and 3×3(blue).

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[2] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019. 2

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1

[4] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed

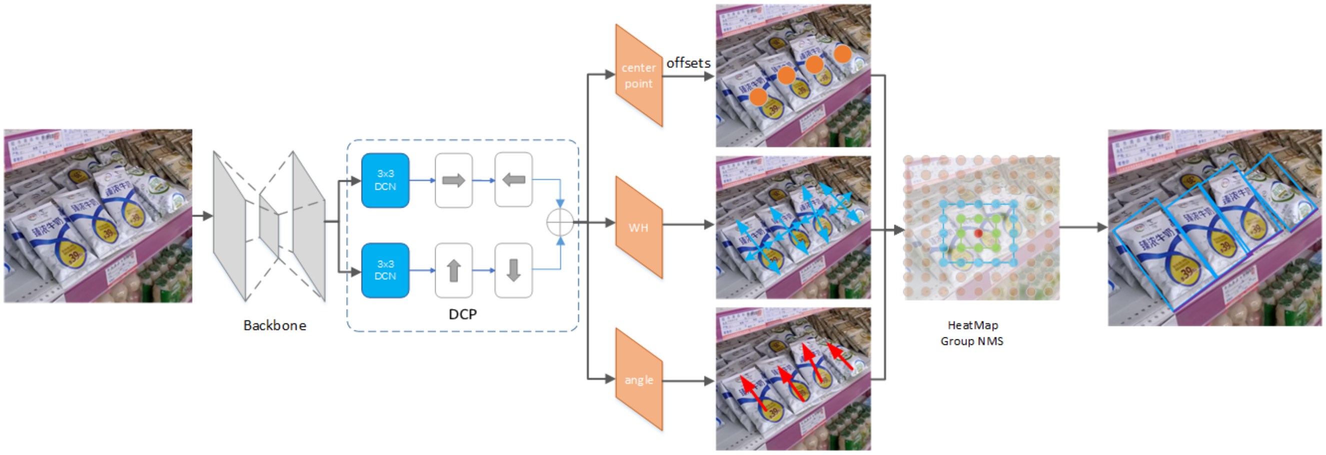
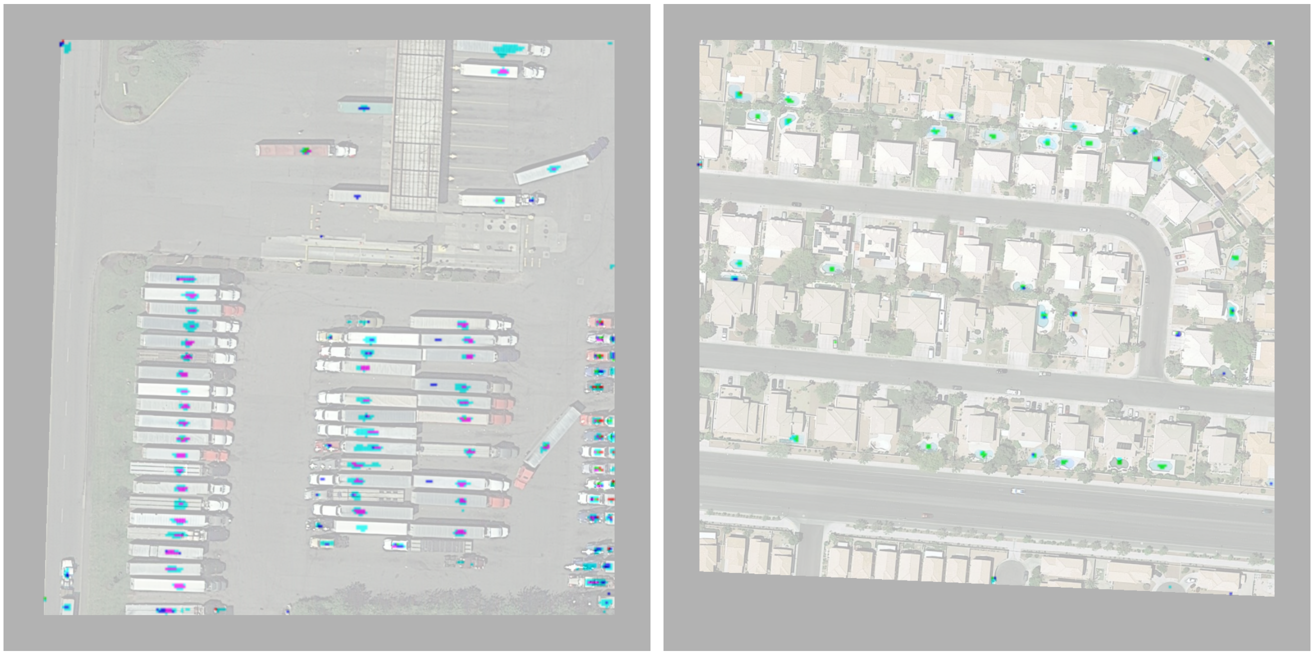


Figure 3. The overall framework of CenterNet†. We insert deformable center pooling layer between backbone and heads.



(a)

(b)

Figure 4. Illustrations of attention maps in FSM.

scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5227–5236, 2019. 1

[5] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 1

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1