A Robust Detector for Pedestrian Detection

Hongkai Zhang, Bingpeng Ma, Hong Chang, Shiguang Shan, Xilin Chen
{Hongkai.Zhang, Bingpeng.Ma, Hong.Chang, Shiguang.Shan, Xilin.Chen}@vipl.ict.ac.cn
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

Introduction

We proposes a robust detector especially for pedestrian detection. Considering the diversity of the dataset, we use a series of powerful methods to enhance the robustness of our model. Since there exists various sizes of persons in the images, we adopt FPN[1] to extract features from different levels. Due to the mAP evaluation metric, localization performance is very important, so we perform cascaded detection like Cascade R-CNN[2]. For better feature extraction, we use some useful modules like Deformable Convolution[3], Re-weighting pooled features[4], ROI-Align[5], etc. Moreover, context information is encoded in the features for occlusion handling. For robustness, we use data augmentations like changing gamma, changing saturation, gaussian blur and random cropping. Multi-scale testing and ensemble are used for better results.

Data Analysis

Since data is extremely important to the network's performance, we first did data analysis. There are several cases which can be seen from *Fig 1*:

- Different brightness and scenes
- Various sizes of persons
- Occlusion
- Unlabelled persons

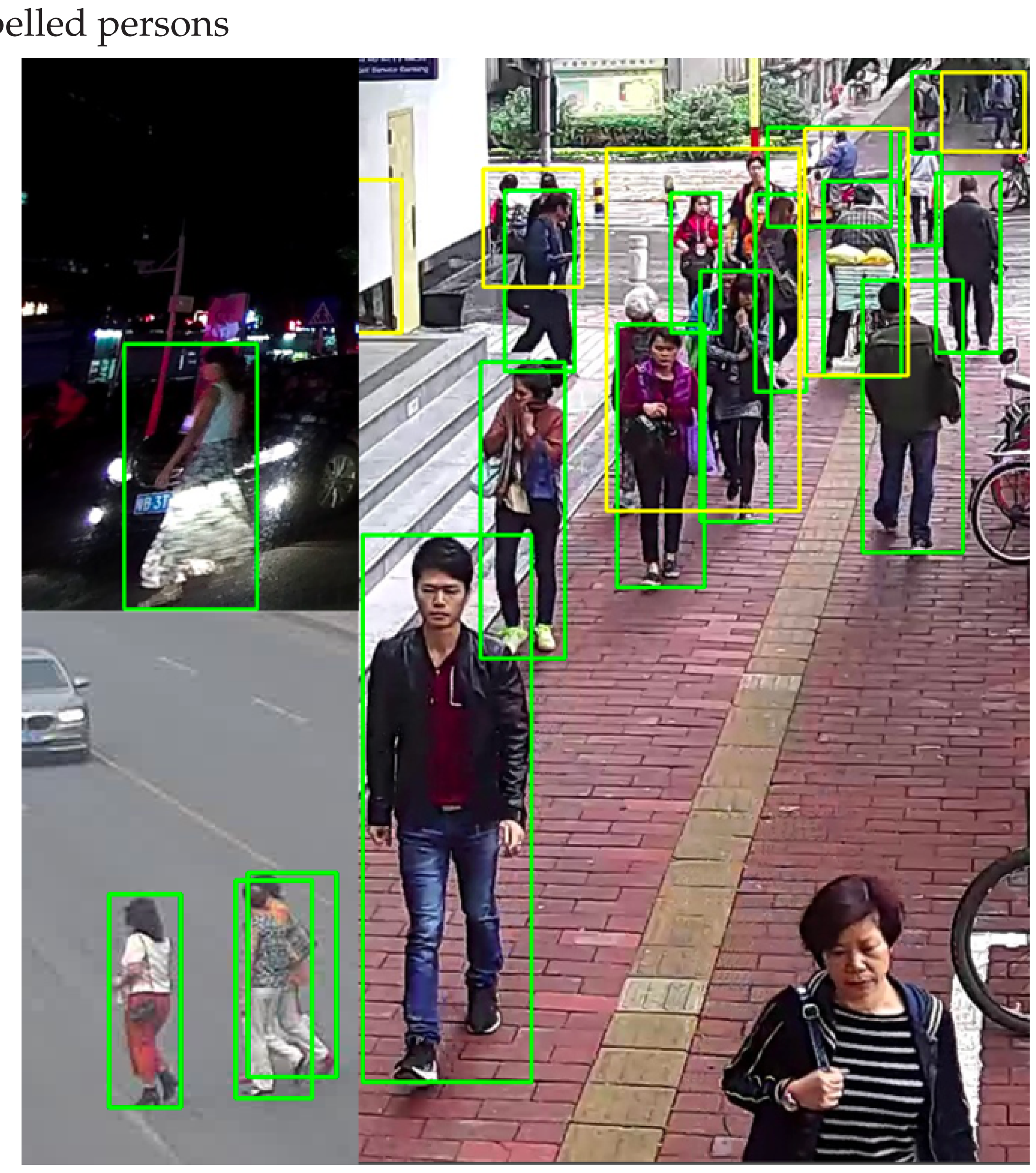

Fig 1: Wider Pedestrian Data Analysis.

In the image, green boxes represent ground-truth boxes, yellow boxes are ignored boxes. We are able to find that the woman in the lower right corner is not labelled, and the two persons in the lower left corner have a big overlap. And it is obviously that there are different scenes and various sizes of persons. These are the challenges of this dataset.

Training and Testing

Training:

- Data augmentation: change gamma, saturation, gaussian blur, random crop, etc.
- Multi label: regarding pedestrian and cyclist as different labels.

Testing:

- Multi scale testing(4 scales with flipped): we merge the results from different scales, adopt soft-nms[7] and do box-voting.
- Ensemble: we split the network into RPN-net and RCNN-net, select proposals from the result of all RPN-nets, send them into RCNN-net to get the results form different models, then normalize scores and coordinates.

References

[1] Tsung-Yi Lin, et al. Feature Pyramid Networks for Object Detection, CVPR2017.
[2] Zhaowei Cai, et al. Cascade R-CNN: Delving Into High Quality Object Detection, CVPR2018.
[3] Jifeng Dai, et al. Deformable Convolutional Networks, ICCV2017.
[4] Shanshan Zhang, et al. Occluded Pedestrian Detection Through Guided Attention in CNNs, CVPR2018.
[5] Kaiming He, et al. Mask R-CNN, ICCV2017.
[6] Kaiming He, et al. Deep Residual Learning for Image Recognition, CVPR2016.
[7] Navaneeth Bodla, et al. Soft-NMS – Improving Object Detection With One Line of Code, ICCV2017.
[8] Shuai shao, et al. CrowdHuman: A Benchmark for Detecting Human in a Crowd, Arxiv2018.
[9] Jie Hu, et al. Squeeze-and-Excitation Networks, CVPR 18.

Architecture

Overall Architecture:

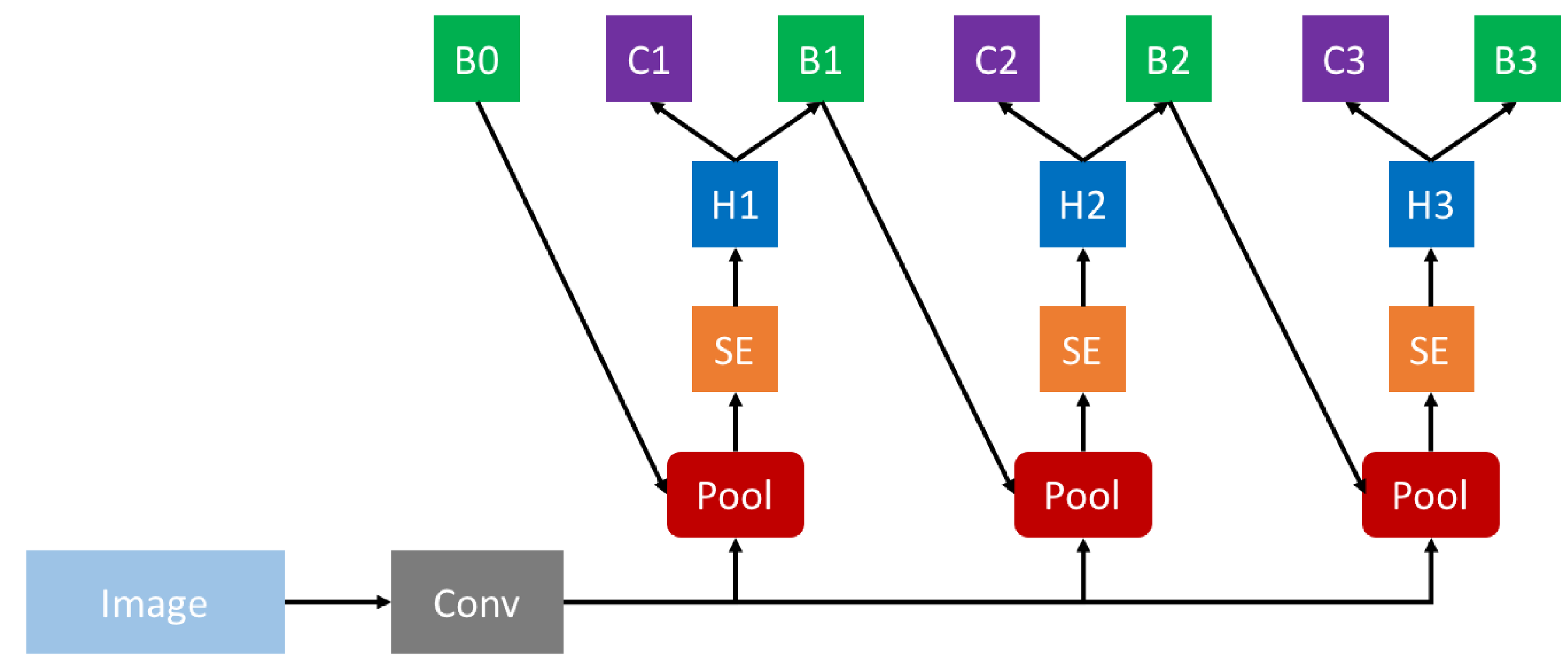

Fig 2: Overall architecture.

Basemodel:

- We use Resnet-50[6] since it's a very powerful and popular basemodel
- Deformable Convolution[3] is used for better feature extraction, especially for occlusion handling

FPN with Cascade R-CNN:

- FPN[1] is adopted to handle different scales of person
- Cascade R-CNN[2] helps to achieve more accurate localization performance, since localization is very important in the AP metric

Useful modules:

- ROI-align[5] uses bilinear interpolation instead of quantization when pooling features, this get more precise features
- Re-weight Pool5[4] adds an channel-wise attention after pooled features to focus on important channels. It is useful for occlusion handling
- Context information: We concatenate the FCs for better classification performance as in *Fig 3*

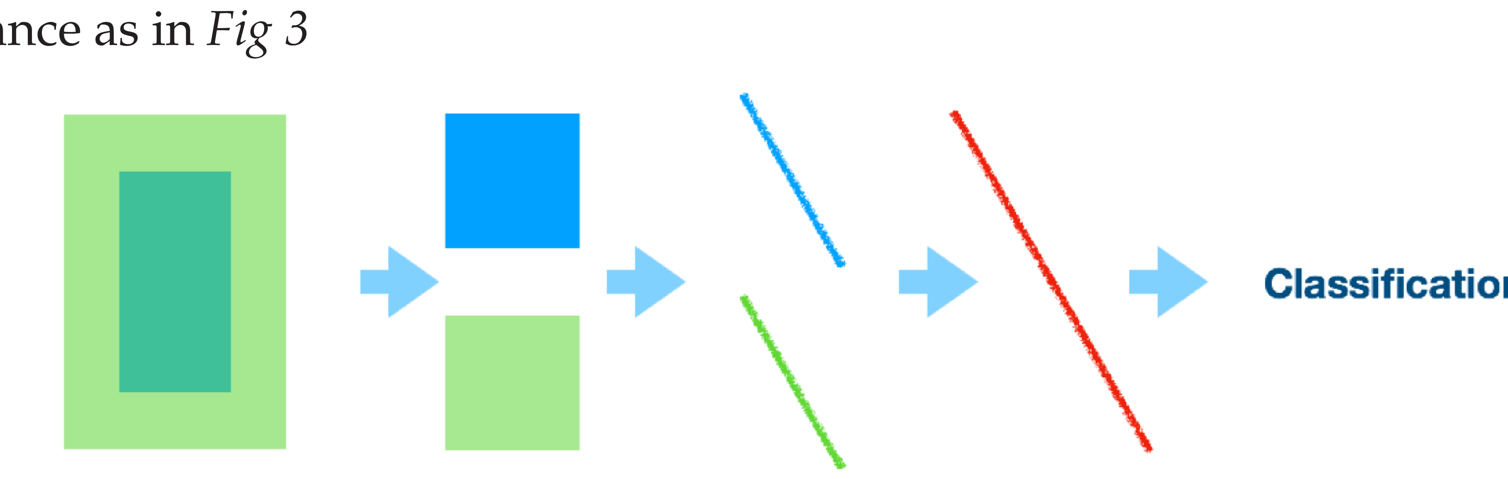

Fig 3: Context information.

Results

Method	*Comments*	*AP*
Cascade RCNN	3 stage [0.5, 0.6, 0.7]	+3.8
Deformable conv	-	+0.8
Reweight Pool5	-	+0.8
Multi label	specify person and cyclist	+0.4
Augmentation	color and random crop	+3.5
Bn training	-	+1.3
Multi-scale testing	4 scale with flip	+2.9
Ensemble	4 models	+2.2
Single Res-50	1st submission, full image training	63.21
Single SE-152[9]	2nd submission, full image training	66.57
3 model Ensemble	3rd submission, random crop training	68.72
5 model Ensemble	4th submission, random crop training	69.56
6 model Ensemble	add a Crowd Human[8] pre-trained model	69.63
JDAI-Human	-	64.4
NtechLab	-	62.49