

PCN: Part and Context Information for Pedestrian Detection with CNNs

Shiguang Wang¹
xiaohu_wyyx@163.com

Jian Cheng¹
chengjian@uestc.edu.cn

Haijun Liu¹
haijun_liu@126.com

Ming Tang²
s1646687@sms.ed.ac.uk

¹ School of Electronic Engineering,
University of Electronic Science and
Technology of China

² Science&Engineering,
The University of Edinburgh

Abstract

Pedestrian detection has achieved great improvements in recent years, while complex occlusion handling is still one of the most important problems. To take advantage of the body parts and context information for pedestrian detection, we propose the part and context network (PCN) in this work. PCN specially utilizes two branches which detect the pedestrians through body parts semantic and context information, respectively. In the Part Branch, the semantic information of body parts can communicate with each other via recurrent neural networks. In the Context Branch, we adopt a local competition mechanism for adaptive context scale selection. By combining the outputs of all branches, we develop a strong complementary pedestrian detector with a lower miss rate and better localization accuracy, especially for occlusion pedestrian. Comprehensive evaluations on two challenging pedestrian detection datasets (i.e. Caltech and INRIA) well demonstrated the effectiveness of the proposed PCN.

1 Introduction

Pedestrian detection, a canonical sub-problem in computer vision, has been extensively studied in recent years [6, 15, 19, 22, 28, 31, 32]. It has many applications such as video surveillance [3] and intelligent vehicles [9]. Although pedestrian detection has achieved steady improvements over the last decade, complex occlusion is still one of the obstacles. According to the statistic analysis in [9], over 70% of the pedestrians are occluded at least in one video frame. For example, the current best-performing detector RPN+BF [31] attained 7.7% of the average miss rate on Caltech [9] under the none occlusion setting. However, when heavy occlusions are present, its performance is much worse than several state-of-the-art algorithms. The detailed results are shown in Table 2.

Recently, DeepParts [22] was specially designed to cope with the occlusion cases in pedestrian detection, by constructing an extensive part pool for automatic parts selection. The well-designed parts are selected driven by data, requiring less prior knowledge of the occlusion types than [20, 30]. However, the semantic parts cannot communicate with each

other, which is very important for occluded pedestrian handling. As illustrated in Fig. 1, when left hand is occluded, the corresponding part score may be very low. Consequently, the summed score will also be low. The final result is failed to recognize the occluded pedestrian. However, with semantic communication, the visible parts such as head and middle body can pass messages to support the existence of left hand. Thus, the body part scores could be refined by this semantic information communication and we are more likely to recognize the occluded pedestrian successfully.

Inspired by ION [20], we introduce LSTM [26] for semantic parts communication due to its extraordinary ability for learning long-term dependences. As shown in Fig. 1, the pedestrian box is divided into several part grids. Each part grid corresponds to a detection score. With LSTM, those body parts could memorize the semantic information and communicate with other parts to refine the score map. Moreover, different body parts need different information from other semantic parts, therefore, we use the gates function in LSTM to control the message passing.

On the other hand, a lot of researches [4, 6, 10] stressed that context information plays an important role in object detection. However, they usually only extracted the context information in a single scale, where the scale is set by handcraft [6, 10]. Considering that different pedestrian instance may need different context information, we propose to use context regions with different scales to extract the context information. To adaptively select the context information in different scales, a local competition operation (maxout [14]) is introduced into our model. For example, the small size pedestrian may need more context information due to low resolution, whilst the heavy-occluded pedestrian may need more context information than the no-occluded pedestrian.

Contributions. To sum up, this work makes the following contributions. (1) we propose a novel part and context network (PCN) for pedestrian detection, which incorporates an original branch, a part branch and a context branch into an unified architecture. (2) To the best of our knowledge, we are the first to introduce the LSTM module into pedestrian detection framework for body parts semantic information communication. (3) The context branch can adaptively handle multi-scale context information in a data driven manner via a local competition mechanism. (4) The results on Caltech and INRIA datasets demonstrated the effectiveness of our proposed model PCN for pedestrian detection with a lower miss rate and a better localization accuracy, especially for those occluded pedestrians.

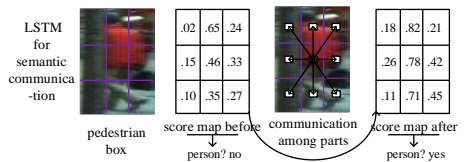


Figure 1: Motivation of our part branch. It’s hard to classify the pedestrian box because of heavy occlusion. With communication among parts, the visible parts and invisible parts could pass message bi-directionally and complement each other to support the existence of pedestrian. Thus, the refined body parts score map is prone to classify the pedestrian box.

2 Related Work

In this section, we mainly review related works in three aspects.

Part-based pedestrian detectors. Part-based approaches for pedestrian detection can be classified into two categories: supervised manner and unsupervised manner. Mohan et

al. [22] and Krystian et al. [21] firstly trained part detectors in a fully supervised manner and then combined their outputs to fit a geometric model. Another category of part-based models focus on unsupervised part mining, which does not require part labels. Wang et al. [29] introduced deformable part model (DPM) to handle the pose variations by learning a mixture of local templates for each body part. Lin et al. [17] proposed a promising framework by incorporating DPM into And-Or graph. However, DPM [29] needs to handle complex configurations. Recent work DeepParts [27] constructed an extensive part pool for automatic parts selection, but its performance was restricted by the handcrafted part pool.

Multi-context features. Multi-context or multi-region features are useful in object detection [2, 8, 11, 23]. Gidaris et al. [11] firstly introduced multi-region into deep convolutional neural network and explored their role in detection. Bell et al [8] used spatial recurrent neural networks to integrate multi-context information of RoI. Ouyang et al. [23] applied RoI-Pooling from image features using different context regions and resolutions.

Deep models. Deep learning methods are widely used in pedestrian detection. For instance, Hosang [15] demonstrated the effectiveness of the R-CNN pipeline [13] in pedestrian detection. Cai et al. learned complexity-aware cascades for deep pedestrian detection [6] and also proposed a unified multi-scale deep CNN for fast object detection [8]. Zhang et al. [30] used a region proposal network (RPN) [24] to compute pedestrian candidates and a cascaded boosted forest [10] to perform sample re-weighting for candidates classification. Du et al. [10] introduced a soft-rejection based network fusion method to fuse the soft metrics from all networks together to generate the final confidence scores, achieving state-of-the-art results on Caltech pedestrian dataset [9].

3 Our Approach

To take advantage of the body parts and context information for pedestrian detection, we propose the part and context network (PCN). PCN specially utilizes two sub-networks which detect the pedestrians through body parts semantic information and context information, respectively.

3.1 Architecture

The detailed architecture of our pedestrian detector is shown in Fig. 2. Based on the framework of Faster RCNN [24], our network detects objects through 3 detection branches: the original branch, the part branch and the context branch. Each branch has different focus. The original branch uses original box from RPN [24] so that it can focus on the full body of pedestrian. The part branch is designed to make full use of semantic parts information for precise classification, specially for pedestrians with heavy occlusion. Considering that pedestrian instances may have different occlusion states, part score maps related to pedestrian box are modelled as sequence problem to communicate with each other. In the part branch, pedestrian box is divided into several part grids (e.g., 3×3) after RoI-Pooling and several convolution layers, where each part grid is associated with a detection score. Those semantic scores are applied for pedestrian parts communication using LSTM. Since multi-region, multi-context features were found to be effective for object detection [2, 11, 23], a context branch is introduced to obtain context features with diversity among multiple scales. Under the motivation that different pedestrian instances may need different context information, we adopt a local competition mechanism (maxout) among multiple scales to increase

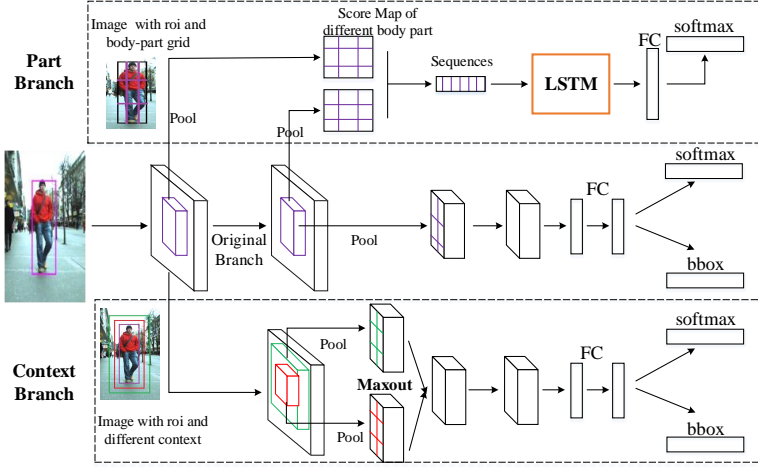


Figure 2: Architecture of our model. The unified network consists of 3 subnetworks: a part branch using LSTM for semantic information communication, an original branch using original box from RPN, a context branch introducing maxout operation for region scale selection.

the adaptive ability of context selection. The results produced by all detection branches are weighted sum as the final detection scores. In our implementation, “ \hat{a} trous” [18] trick is used to increase the feature map resolution.

3.2 LSTM for part semantic information communication

Our architecture for part semantic features communication in PCN is shown in detail in Fig. 3. To get ROI features with different resolution, we place RoI-Pooling layer on top of different conv layers. Features of *conv3_3* and *conv4_3* achieved best performance in our experiments. $L2$ -normalization is used following [18] to make training more stable. To get a parts score map of the pedestrian instance, the ROI feature maps are passed through several convolution layers to reduce channel dimension and resolution. Each branch gets a parts score map with resolution of $K \times K$ ($K=3$ in Fig. 3) after the softmax layer. The score maps have strong semantic information of pedestrian, for example the top-left grid indicates the left-head-shoulder part. These parts score maps are concatenated for communication in the next step.

To model coherence between semantic parts, we introduce LSTM for encoding their relationships due to its long short term memory ability to model sequential data. We first permute, flip, and reshape the parts score map to generate 4 sequences with different orders. Then, the LSTM moves on parts score map along four directions: left, right, top and bottom to encode parts and make semantic communication among them. As demonstrated in Fig. 1 in Section 1, when left hand is occluded, the corresponding part score may be very

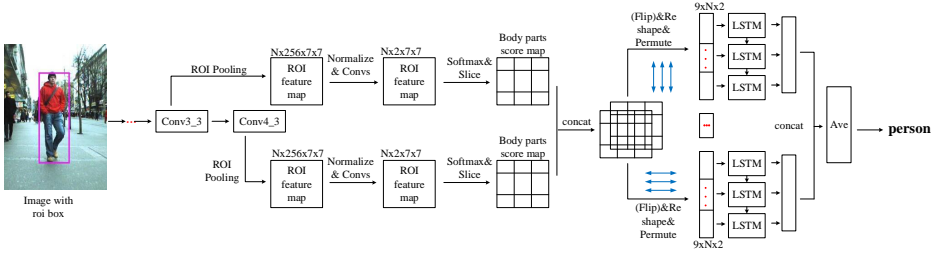


Figure 3: LSTM for part semantic communication. In a single pass, we extract ROI feature from *conv3_3* and *conv4_3*. Each descriptor is $L2$ -normalized. We get parts score maps with size of 3×3 through several conv layers. The body parts score maps after softmax are concatenated, (flipped or not), reshaped, permuted and then sent to LSTM networks. LSTM moves along four directions: left, right, top and bottom. Refined scores outputted from LSTM are averaged to get the final scores of Part Branch.

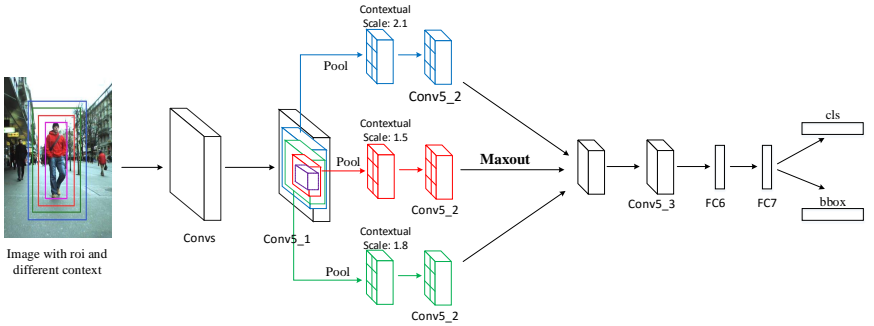


Figure 4: Maxout for adaptive context selection. As shown above, the purple box indicates the original box of pedestrian, the red, green and blue box indicate the original context box with a scale factor of 1.5, 1.8 and 2.0, respectively. After a convolution layer, each context box is sent to maxout for competition and adaptive context selection. Class confidence and bounding box are outputted as the result.

low. However, the visible part such as head and middle body can pass message to support the existence of the left hand. Consequently, the body part scores could be refined by this semantic information communication. The gate functions in LSTM are used to control the message passing.

3.3 Maxout for adaptive context selection

Our architecture for adaptive context selection in PCN is detailedly shown in Fig. 4. Multi-region, multi-context features were found to be effective in [8, 9, 10]. However, they only extracted the context information in a single scale, where the scale is set by handcraft [8, 10]. Considering that different pedestrian instance may need different context information, we propose to use context regions with different scales to extract the context information. For example, small size pedestrian may need more context information due to low resolution, and heavy-occluded pedestrian may need more context information than no-occluded pedes-

trian. Similar philosophy was also figured out in generic object detection [23], which used a chaining structure for multi-context mining. Our context convolutional feature maps are extracted from multiple discrete scales, known as multi-context. In Fig. 4, 3 kinds of region scale are enumerated for generating context box. Suppose the RoI has width W , height H , and S is the context scale factor of the RoI. Thus, the context region has width $W \times S$ and height $H \times S$ with the same center as the original RoI, where $S = 1.5, 1.8, 2.1$ respectively for each context branch in our implementation.

We incorporate a local competition operation (maxout) into PCN to improve the adaptive context selection capability for various context features sources. Each sub-branch with different context scale passes through RoI-Pooling layer to get a fixed-resolution ($m \times m$) object feature map. These feature maps of different data source are selected by maxout with a data driven manner. Maxout [24] (element-wise max) is a widely considered operation for merging two or multiple competing sources. Here, we follow the well-known structure of NoCs [25]. When the maxout operation is used, the three feature maps (for the three context scales) are merged into a single feature map with the same dimensionality. The convolution layers before each RoI-Pooling layer share their weights. Thereby, the total number of weights is almost unchanged when using maxout.

3.4 Training Procedure

Since these branches differ a lot from feature extraction and expression, which could lead to some instability during learning, we train our model by multiple stages.

Stage1: Jointly optimizing original branch and context branch. Original branch and context branch could be jointly optimized mainly because they have the similar feature expression (different from part branch). Therefore, the gradients from the 2 branches could be merged to the trunk in a stable way. The parameters W of original branch and context branch are learned from a set of training samples. Suppose X_i is a training image patch, and $Y_i = (y_i, t_i)$ the combination of its class label $y_i \in \{0, 1, 2, \dots, K\}$ and bounding box coordinates $t_i = (t_i^x, t_i^y, t_i^w, t_i^h)$, we optimize W with a multi-task loss:

$$L(W) = \sum_{m=1}^M \alpha_m l^m(X_i, Y_i | W), \quad (1)$$

where M is the number of detection branches, and α_m is the weight of loss l^m . l^m is a multi-task loss combining the cross-entropy loss and the smoothed bounding box regression [26, 27]. The optimal parameters $W^* = \arg \min_W L(W)$ are learned by stochastic gradient descent.

Stage2: Pre-training the part branch to get parts score maps by the cross-entropy loss $L_{cls}(p(X), y) = -\log p_y(X)$, where $p(X) = (p_0(X), \dots, p_K(X))$ is the probability distribution over classes. Part branch is pre-trained so that parts can focus on information communication in Stage3. The model trained in Stage1 is fixed in Stage2 to save compute cost.

Stage3: Training the part branch for parts semantic information communication using LSTM. Parameters trained in stage1 are fixed to save compute cost, while parameters trained in Stage2 are finetuned in Stage3.

Settings	Description	Settings	Description
Reasonable	50+ pixels, Occ. none or partial	Occ.partial	1-35% occluded
All	20+ pixels, Occ. none, partial or heavy	Occ.heavy	35-80% occluded
Occ.none	0% occluded	Over.75	Reasonable, using IOU=0.75

Table 1: Evaluation settings for Caltech pedestrian dataset [9].

4 Experiments

We evaluated the effectiveness of the proposed PCN on 2 popular pedestrian detection datasets including Caltech [9], INRIA [8]. More experimental analyses on the effectiveness of each component in our network are further given on the challenging Caltech dataset [9].

4.1 Datasets

Caltech. The Caltech dataset and its associated benchmark [9] are among the most popular pedestrian detection datasets. It consists of about 10 hours videos (640×480) collected from a vehicle driving through regular urban traffic. The annotation includes temporal correspondence between bounding boxes and detailed occlusion labels. We use dense sampling of the training data by 10 folds as adopted in [15]. 4024 images in the standard test set are used for evaluation. The miss rate (MR) is used as the performance evaluation metric [9]. The detailed evaluation settings following [9] are shown in Table 1.

INRIA. The INRIA pedestrian dataset [8], which is often used for verifying the generalization capability of models, is split into a training and a testing set. The training set consists of 614 positive images and 1,218 negative images. The testing set consists of 288 testing images. Our model is evaluated on the testing set by MR.

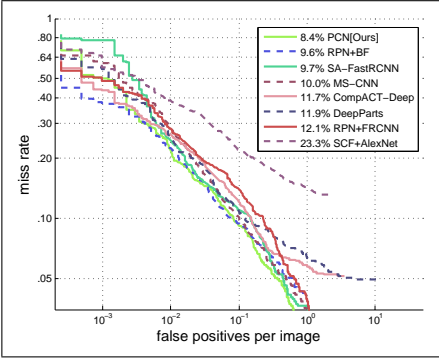
4.2 Implementation Details

For RPN, we used anchors of 9 different scales, starting from 40 pixels height with a scaling stride of $1.4 \times$. Other hyper-parameters of RPN followed [12, 24, 8]. When training, 70k iterations were run with an initial learning rate of 0.001, which decays 10 times after 50k iterations. Model was finetuned from 16-layers VGG-Net. When testing, the scale of the input image was set as 720 and 560 pixels on the shortest side on Caltech and INRIA datasets respectively, we just selected top 50 boxes for proposals.

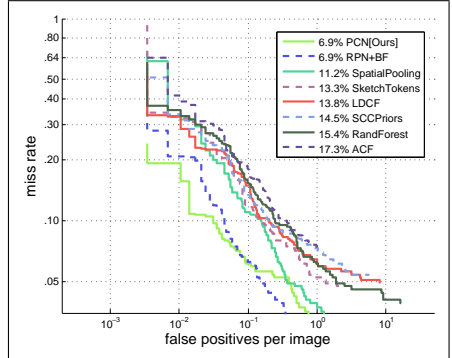
For PCN, we used the hyper-parameters following [16] to finetune our model from 16-layers VGG-Net. The fourth max pooling layer was removed to produce larger feature maps in all branches. “ \hat{a} trous” [9] trick was used to increase the feature map resolution and reduce stride. Based on these settings, the parameter optimized model was used as our basic model (RPN+FCNN_{opt}). When training, top 1000 boxes were selected as proposals, and 50k iterations were run with an initial learning rate of 0.001 for 3 stages, which decays 10 times after 40k iterations. When testing, with the top 50 boxes from RPN, we used nms with 0.5 after weight combined from different branches. Conv5’s channels are reduced to 256, while fc6 and fc7 are reduced to 2048, 512 respectively to reduce computation and memory burden.

Methods	Reasonable	All	Occ.none	Occ.partial	Occ.heavy	Over.75
SCF+AlexNet [15]	23.3	70.3	20.0	48.5	74.7	58.9
DeepParts [27]	11.9	64.8	10.6	19.9	60.4	56.8
CompACT-Deep [8]	11.8	64.4	9.6	25.1	65.8	53.3
SAF R-CNN [16]	9.7	62.6	7.7	24.8	64.4	41.0
MS-CNN [6]	10.0	61.0	8.2	19.2	59.9	56.9
RPN+BF [50]	9.6	64.7	7.7	24.2	69.9	35.5
RPN+FRCNN _{opt} [46]	12.1	65.4	10.5	24.0	64.9	48.5
PCN(Ours)	8.4	61.8	7.0	16.4	56.7	34.8

Table 2: Detailed breakdown performance comparisons of our models and other state-of-the-art models on the 6 evaluation settings.



(a)



(b)

Figure 5: The results on two Datasets under Reasonable evaluation (legends indicate MR) (a) Comparisons on the **Caltech** set; (b) Comparisons on the **INRIA** set.

4.3 Comparisons with State-of-the-art Methods

Caltech. Fig. 5. (a) and Table 2 show the results on Caltech (lower is better). We compare our framework with several other state-of-the-art approaches [8, 6, 15, 16, 24, 27, 50]. Compared with these approaches, our PCN obtained a miss rate of 8.4% on reasonable setting, which is over 1.2 points better than the closest competitor (9.6% of RPN+BF [50]). In the partial and heavy occlusion settings, our PCN achieved at least 3 points improvements than DeepParts [27] which is specially designed for occlusion, demonstrating the effectiveness of our PCN model for occlusion handling. Visualization of detection results by state-of-the-art methods and our PCN model are shown in Fig. 6.

INRIA. Fig. 5. (b) shows the detection results on the INRIA datasets. Our PCN model achieved a miss rate of 6.9%, comparable with the best available competitor RPN+BF [50].

4.4 Ablation Experiments

In order to figure out which component is working, we carried out the following ablation experiments further.

A. Part branch In this subsection, we explore the effect of part branch. In PCN, part branch divides the pedestrian box into several parts. Each part corresponds to multiple part detector (RoI-Pooling from different layers). LSTM is introduced to encode the coherence between parts and make semantic communication among them, which is crucial for occluded pedes-



Figure 6: Comparison of pedestrian detection results with other state-of-the-art methods. The first column shows the input images with ground-truths annotated with red rectangles. The rest columns show the detection results (green rectangles) of DeepParts [27], RPN+BF [51] and PCN respectively. Our PCN can successfully detect more partial or heavy occluded instances with better localization accuracy compared to the other two state-of-the-art methods.

trian detection. To verify this, we first added part detectors on the parameter optimized basic model RPN+RFCNN_{opt} , the detector scores of all parts were averaged for the final score of part branch. As we can see, Base+Part_{avg} has better performance than the Base method, especially the performance of Occ.partial (24.0% vs 21.0%). However, the performance of Occ.heavy increased a little, which indicates that simply splitting pedestrian box to body parts to design part detector is not enough to recognize heavy occluded pedestrian. When introducing LSTM for encoding the coherence between parts (Base+Part+LSTM), the performance of Occ.partial and Occ.heavy improved from 21.0%/64.6% to 17.0%/58.9%, respectively, showing that the communication of semantic parts is very important for pedestrian detection, especially for these heavy occluded instances.

B. Context branch In this subsection, we explore the effect of context branch. To reveal the importance of context information in pedestrian detection, we added a context branch based on the basic model. One scale context only uses the context information generated by window size of $\text{WS} \times \text{HS}$, Table 3 shows the performance comparison when choosing different context scales ($\text{Base+Context}(S=x)$), obtaining a MR of 9.8%/9.9%/9.6% for $x=1.5/1.8/2.1$, respectively. This demonstrated that context plays a key role in pedestrian detection. Interesting, the performance using an IoU threshold of 0.75 ($\text{Over}.75$) increased a lot, the improvement of localization accuracy may be caused by the context information and extra bounding box regulator. For adaptive context selection, we introduced Maxout to integrate context scale. $\text{Base+Context}(\text{Maxout})$ achieved a MR of 9.0%, better than those models with single context scale. Because of length limited, we just listed 3 scales ($x=1.5, 1.8, 2.1$)

Methods	Reasonable	All	Occ.none	Occ.partial	Occ.heavy	Over.75
RPN+FRCNN _{opt} (Base)	12.1	65.4	10.5	24.0	64.9	48.5
Base+Part _{avg}	10.1	62.8	8.1	21.0	64.6	39.0
Base+Part+LSTM	9.4	62.6	8.1	17.0	58.9	38.5
Base+Context(S=1.5)	9.8	63.5	8.1	19.4	60.1	38.0
Base+Context(S=1.8)	9.9	63.4	8.3	20.6	59.8	36.5
Base+Context(S=2.1)	9.6	63.5	8.2	17.8	59.9	36.1
Base+Context(Maxout)	9.0	62.6	7.5	17.5	57.5	34.6
PCN(full)	8.4	61.8	7.0	16.4	56.7	34.8

Table 3: Detailed breakdown performance comparisons of ablation experiments: the effect of part branch and context branch. RPN+FRCNN_{opt}: the parameter optimized RPN+FRCNN; Base+Part_{avg}: basic model add part detectors; Base+Part+LSTM: basic model add part detectors and semantic parts communication; Base+Context(S=x): basic model add context with single scale x (x=1.5,1.8,2.1); Base+Context(Maxout): basic model add context using Maxout for adaptively scale selection.

for illustration.

5 Conclusion

In this paper, we proposed the part and context network (PCN). PCN specially utilizes two sub-networks which detect the pedestrians through body parts semantic information and context information, respectively. Extensive experiments demonstrated that the proposed PCN is superior in detecting occluded pedestrian instances and achieving better localization accuracy.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (61671125, 61201271, 61301269), and the State Key Laboratory of Synthetical Automation for Process Industries (NO. PAL-N201401).

References

- [1] Ron Appel, Thomas J Fuchs, Piotr Dollár, and Pietro Perona. Quickly boosting decision trees-pruning underachieving features early. In *ICML (3)*, pages 594–602, 2013.
- [2] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2874–2883, 2016.
- [3] Muhammad Bilal, Asim Khan, Muhammad Umar Karim Khan, and Chong-Min Kyung. A low complexity pedestrian detection framework for smart video surveillance systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

- [4] Alberto Broggi, Alex Zelinsky, Ümit Özgüner, and Christian Laugier. Intelligent vehicles. In *Springer Handbook of Robotics*, pages 1627–1656. Springer, 2016.
- [5] Zhaowei Cai, Mohammad Saberian, and Nuno Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3361–3369, 2015.
- [6] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision*, pages 354–370. Springer, 2016.
- [7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.
- [8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.
- [9] P Dollár, C Wojek, B Schiele, and P Perona. Pedestrian detection: an evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012.
- [10] Xianzhi Du, Mostafa El-Khamy, Jungwon Lee, and Larry S. Davis. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. *CoRR*, abs/1610.03466, 2016. URL <http://arxiv.org/abs/1610.03466>.
- [11] Spyros Gidaris and Nikos Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1134–1142, 2015.
- [12] Ross Girshick. Fast r-cnn. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.
- [13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [14] Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron C Courville, and Yoshua Bengio. Maxout networks. *ICML (3)*, 28:1319–1327, 2013.
- [15] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. Taking a deeper look at pedestrians. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4073–4082, 2015.
- [16] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. *arXiv preprint arXiv:1510.08160*, 2015.
- [17] L. Lin, X. Wang, W. Yang, and J. H. Lai. Discriminatively trained and-or graph models for object shape detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):959–72, 2015.

- [18] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.
- [19] Ping Luo, Yonglong Tian, Xiaogang Wang, and Xiaoou Tang. Switchable deep network for pedestrian detection. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 899–906, 2014.
- [20] Markus Mathias, Rodrigo Benenson, Radu Timofte, and Luc Van Gool. Handling occlusions with franken-classifiers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1505–1512, 2013.
- [21] Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*, pages 69–82. Springer, 2004.
- [22] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.
- [23] Wanli Ouyang, Ku Wang, Xin Zhu, and Xiaogang Wang. Learning chained deep features and classifiers for cascade in object detection. *arXiv preprint arXiv:1702.07054*, 2017.
- [24] S. Ren, K. He, R Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis Machine Intelligence*, pages 1–1, 2016.
- [25] Shaoqing Ren, Kaiming He, Ross Girshick, Xiangyu Zhang, and Jian Sun. Object detection networks on convolutional feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 50(1):815–830, 2015.
- [26] Lambert M. Surhone, Mariam T. Tennoe, and Susan F. Henssonow. Long short term memory. *Betascript Publishing*, 2010.
- [27] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning strong parts for pedestrian detection. In *IEEE International Conference on Computer Vision*, pages 1904–1912, 2015.
- [28] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian detection aided by deep learning semantic tasks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5079–5087, 2015.
- [29] Xiaogang Wang and Wanli Ouyang. A discriminative deep model for pedestrian detection with occlusion handling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3258–3265, 2012.
- [30] C Wojek, S Walk, S Roth, and B Schiele. Monocular 3d scene understanding with explicit occlusion reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, Co, Usa, 20-25 June*, pages 1993–2000, 2011.

- [31] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *European Conference on Computer Vision*, pages 443–457. Springer, 2016.
- [32] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1259–1267, 2016.