Density-Aware Pedestrian Proposal Networks for Robust People Detection in Crowded Scenes

Sangdoo Yun, Kimin Yun, Jongwon Choi, and Jin Young Choi(✉)

ASRI, Department of Electrical and Computer Engineering,
Seoul National University, Seoul, South Korea
{yunsd101,ykmwww,jychoi}@snu.ac.kr, jwchoi.pil@gmail.com

Abstract. In this paper, we propose a density-aware pedestrian proposal network (DAPPN) for robust people detection in crowded scenes. Conventional pedestrian detectors and object proposal algorithms easily fail to find people in crowded scenes because of severe occlusions among people. Our method utilizes a crowd density map to resolve the occlusion problem. The proposed network is composed of two networks: the proposal network and the selection network. First, the proposal network predicts the initial pedestrian detection proposals and the crowd density map. After that, the selection network selectively picks the final proposals by considering the initial proposals and the crowd density. To validate the performance of the proposed method, experiments are conducted on crowd-scene datasets: WorldExpo10 and PETS2009. The experimental results show that our method outperforms the conventional method and achieves near real-time speed on a GPU (25 fps).

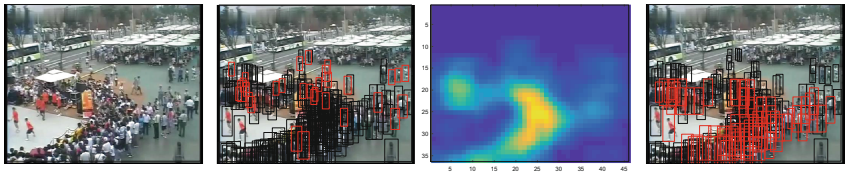
1 Introduction

1.1 Motivations and Objectives

Pedestrian detection is one of the most important tasks in computer vision. Recently, deep convolutional neural network (DCNN) based pedestrian detectors [7,10,14,18] have achieved state-of-the-art performance. DCNN-based detection systems mainly have two stages: First, numerous object candidates, often called "proposals," are extracted as preprocessing through regression on possible locations of pedestrians. Second, DCNN-based detector classifies the proposals and determines the final detections. Since regions except the proposals in image are ignored, high-quality proposals could improve detection performance by rejecting false positives and make whole detection speed faster than using the traditional *sliding window* methods. Therefore, generation of high-quality proposals is essential for fast and accurate object detection.

In this work, we focus on the challenging crowded scene as shown in Fig. 1. Our goal is to find high-quality pedestrian proposals in highly crowded scenes. The main difficulty of generating pedestrian proposals in a crowded scene is caused by severe occlusions among people. Commonly used pedestrian proposal methods, such as LDCF [11] or ACF [4] detectors, are known to show a weak

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part II, LNCS 9914, pp. 643–654, 2016.
DOI: 10.1007/978-3-319-48881-3_45



(a) (b) (c) (d)

Fig. 1. (a) An example of a crowded scene. **(b)** The results of the LDCF detector [11]. The black boxes are missed pedestrians, and the red boxes are the correctly detected pedestrians (IoU > 0.5). **(c)** The crowd density map of the scene. **(d)** The results of the proposed methods. (Color figure online)

performance in a crowded scene with severe occlusions as shown in Fig. 1, which results in slow speed compared with other fast proposal algorithms [14,21]. However, these methods [14,21] do not consider the occlusion and just try to detect each pedestrian independently. Therefore, understanding the occlusion patterns in the whole scene is the key factor to solve this crowded-scene pedestrian proposal problem.

In this paper, we propose a density-aware pedestrian proposal network (DAPPN) to solve the occlusion problem in a crowded scene. We observe that there exists an intimate relationship between crowd density and people detections. For example, a high-density region is more likely to have occlusions of people than a lower-density region. Following this observation, we propose a pedestrian proposal network that considers crowd density in a global view. The proposed DAPPN is composed of two networks: the *proposal network* and the *selection network*. The *proposal network* predicts a density map in a crowded-scene image in addition to rough pedestrian proposals as bounding boxes by RPN [14]. The coarse proposals from RPN [14] are quite useful but not enough to handle the severe occlusion case. In order to obtain high-quality pedestrian proposals in a crowded scene, we introduce a density-aware selection network. The *selection network* selectively picks the final pedestrian proposals by considering both the coarse proposals and the crowd density from the *proposal network*. Since the structure of the proposed network still forms a feed-forward network, the forward pass takes only 40 ms on a GPU.

In experiments, we evaluate the *recall* performance of our method in the popular crowded-scene datasets: *WorldExpo10* [20] and *PETS2009* [1]. The experiments show that the proposed method outperforms the conventional proposal generation algorithms.

1.2 Related Works

Pedestrian Detection. Traditional pedestrian detectors such as DPM [6], ACF [4], and LDCF [11] utilize low-level features to recognize people in images. As a result of their simple and efficient structure, they have demonstrated promising detection quality with acceptable speed using only CPUs. In recent

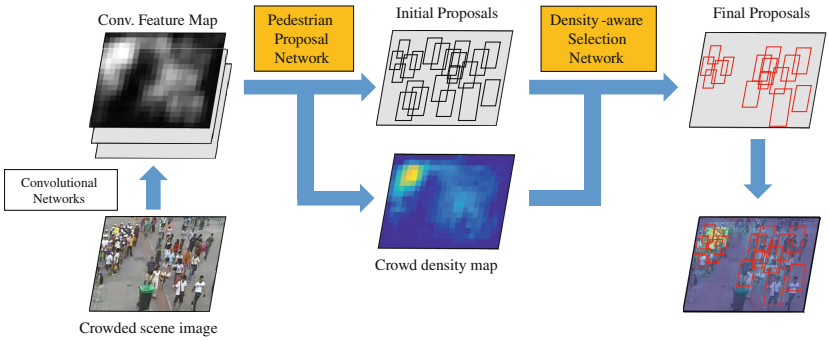


Fig. 2. The framework of the proposed system.

years, DCNN-based detectors [7–9,17] have achieved excellent performance in object classification and detection challenges such as PASCAL VOC [5] and ILSVRC [15]. These DCNN-based detectors strongly require high-quality object proposals as preprocessing. In a crowded scene, Ouyang *et al.* proposed pedestrian detectors handling occlusions by modeling the visibility of individual pedestrian [13] and training occlusion patterns of two coupled people [12]. However, they had difficulty in handling the global occlusion patterns in a crowded scene since they only regard partial occlusions in a local perspective.

Object Proposal. Using the object region proposals in images, object detection algorithms can be faster with significantly reduced search space than using the *sliding-window*-based searching strategy. Also, one can expect a better precision performance by rejecting false positives during the object proposal process than using the sliding window method. There is a wide literature on object proposal methods based on a segmentation approach [2,3,16] and low-level features [21]. The segmentation-based methods promise high-quality object proposals; however, they have a computational bottleneck in object detection because of their time-consuming procedure. EdgeBoxes [21] utilizes edges to provide reasonable proposal quality and fast speed within 1 second. Ren *et al.* proposed a DCNN-based object proposal algorithm, Region Proposal Network(RPN) [14]. It achieved a high detection rate and near real-time speed with the help of GPU computing. However, the above mentioned proposal methods do not consider the severe occlusions in image; therefore, they have difficulty dealing with a crowded scene.

2 Density-Aware Pedestrian Proposal Networks

The overall framework of the proposed system is illustrated in Fig. 2. The proposed system is composed of two networks: the pedestrian proposal network (PPN) and the density-aware selection network (DASN). The convolutional feature map is extracted from the pre-trained convolutional network. The PPN

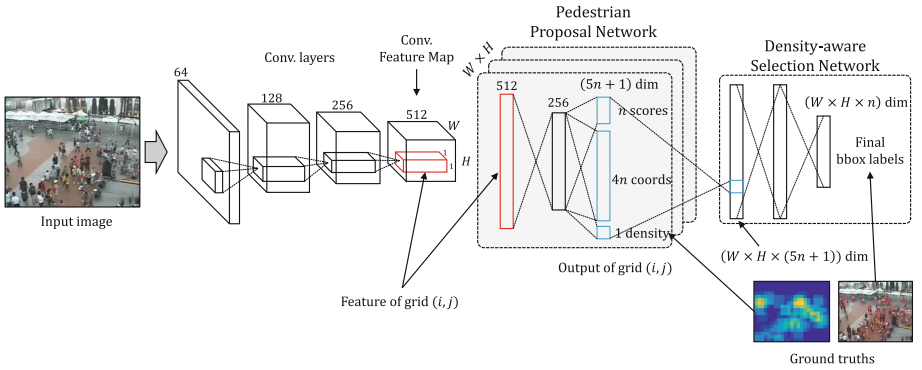


Fig. 3. The proposed density-aware pedestrian proposal network. The network is composed of mainly two parts: pedestrian proposal network, and density-aware selection network. Details of the network are described in Sect. 2.

predicts the initial proposals and the crowd density map. Since the initial proposals from the PPN are independently predicted in local view, the occlusion patterns are not considered. For the global view, the DASN selects the final pedestrian proposals from the initial proposals and the crowd density map. The detailed network structure of the proposed density-aware pedestrian proposal network (DAPPN), is illustrated in Fig. 3.

In Sect. 2.1, we describe the structure of the pedestrian proposal network and the designed multi-task loss function to jointly predict proposals and the density map. In Sect. 2.2, we introduce the density-aware selection network which decides the final pedestrian proposals considering overall scene occlusions.

2.1 Pedestrian Proposal Networks

For the convolutional feature map, we use the 13 layers of VGG-16 model [17] which is pre-trained on ImageNet Dataset [15]. The convolutional feature map has the size $W \times H$ and each cell has 512-dimensional depths. The feature vector of each grid ($1 \times 1 \times 512$) is fed into the fully connected module to predict n proposals and the crowd density of the grid location (n=24 in our experiments). The fully connected module has an intermediate layer to encode features into 256-dimensional vectors. The encoded features are passed through three type of output layers: classification layer to determine the confidence of the proposal, regression layer to localize the proposal's bounding box(bbox), and density regression layer. The classification, bbox regression, and density regression layers output n classification scores, $4n$ bbox coordinates $[x, y, w, h]$, and the scalar value of crowd density at the grid location respectively. Over all the grid locations, the network outputs $(W \times H \times n)$ bounding boxes with scores and crowd density map whose size of $(W \times H)$. The fully connected layers are shared for the features of every grid positions. The shared network reduces a risk of overfitting problem [14].

Anchors. We follow [14] to simultaneously predict multiple pedestrian proposals at each feature location. To regress n bounding boxes at each grid location, each grid has n *anchors* as the reference of proposal bounding boxes. An anchor is represented by the related location (a_x, a_y) from grid center and its size (a_w, a_h). To train the network stably, the anchors should have similar sizes and positions with the true bounding boxes. In order to train anchors, the bounding boxes of training data are grouped into k clusters with their sizes (w, h). In our experiments, we used K-means clustering algorithm ($k = 6$). And also each feature grid is divided into 4 sub-regions, having k anchors in each sub-region. Therefore, the number of predicted bboxes at each grid becomes 24 ($n = 4k$).

Multi-task Loss Function. To train the pedestrian proposal network, we first assign the binary labels (true or false) to the predicted bboxes. For each ground truth box, we select a the predicted box with the highest Intersection-over-Union(IoU) ratio. The selected boxes mean positive samples and the labels are assigned as true. The predicted bboxes who are not selected as corresponding to a ground truth are assigned as negative samples and the labels are assigned as false. If IoU ratio of the selected box is below than 0.5, which means the selected box is not matched well with the ground truth box, then the selected box is treated as negative labels.

Using the assigned labels, we optimize the objective function in order to jointly predict the bounding boxes of proposals and the crowd density map. We formulate the multi-task loss function inspired by the loss function of Fast R-CNN [7]. Letting optimization variables c_i and b_i be the confidence score and bounding box representation for the i-th anchor, and additional variable $\mathbf{d}$ be the vectorized density map, the loss function is defined as follows,

$$\begin{aligned} Loss(\{c_i\}, \{b_i\}, \mathbf{d}) &= \frac{\lambda_c}{N_p} \sum_i L_{conf}(c_i, c_i^*) + \frac{\lambda_b}{N_p} \sum_i c_i^* L_{reg}(b_i, b_i^*) + \lambda_d L_{den}(\mathbf{d}, \mathbf{d}^*), \\ L_{conf}(c_i, c_i^*) &= c_i^* \log \sigma(c_i) + (1 - c_i^*) \log(1 - \sigma(c_i)), \\ L_{reg}(b_i, b_i^*) &= \|b_i - b_i^*\|_1, \\ L_{den}(\mathbf{d}, \mathbf{d}^*) &= \|\mathbf{d} - \mathbf{d}^*\|_2, \end{aligned} \tag{1}$$

where $(\cdot)^*$ indicates the ground truth of $(\cdot)$, L_{conf}, L_{reg}, and L_{den} denote the confidence score loss, bbox regression loss, and density estimation loss respectively. The loss terms are weighted by the parameters λ_c, λ_b, and λ_d, and the number of samples N_p. In the confidence score loss, $\sigma(\cdot)$ denotes the sigmoid function and we use the sigmoid cross entropy loss for binary classification. The bounding box representation (b_i and b_i^*) is expressed by the parameterized coordinates [7]. The regression loss is given by L_1 norm and triggered only for the positive samples ($c_i^* = 1$). The density estimation loss is given by L_2 norm.

2.2 Density-Aware Selection Network

The pedestrian proposals from pedestrian proposal network, called initial proposals, are independently generated from each grid of the convolutional feature

map. Therefore, since they are not considering the whole scene, the initial proposals may contain redundant proposals or miss pedestrians in the occlusion situation. We address this problem by considering the crowd density map. For example, high-density regions have high probability of occlusion, therefore the abundant proposals are needed. Conversely, in low-density area, since the proposals are likely to be false positives, a small number of proposals are necessary.

The proposed density-aware selection network (DASN) generates the final proposals from the initial proposals by awaring the crowd density map. As shown in Fig. 3, the DASN consists of two fully connected layers. The total output of the proposal network for all grids ($W \times H$) are aggregated and vectorized into ($W \times H \times (5n + 1)$) dimensional vector. This vector, which contains the entire proposals (bboxes and scores) and the density map, is fed into the selection network as input. Through the 1024-dim intermediate layer, the output layer predicts ($W \times H \times n$)-dim final bbox confidence scores.

The objective of the selection network is to re-score the confidence scores of proposals in global view. Similarly to the PPN, we assign the positive label ($s_i^* = 1$) to i-th proposal bbox when the proposal is matched with ground truth bbox (IoU > 0.5), otherwise, negative label ($s_i^* = 0$) is assigned. The loss function of DASN is defined as follows,

$$Loss(\{s_i\}) = \frac{1}{N_s} \sum_i \|s_i - s_i^*\|_2, \tag{2}$$

where s_i is the predicted score of i-th proposal and s_i^* is the ground truth label. We empirically select L_2 norm to optimize the selection network.

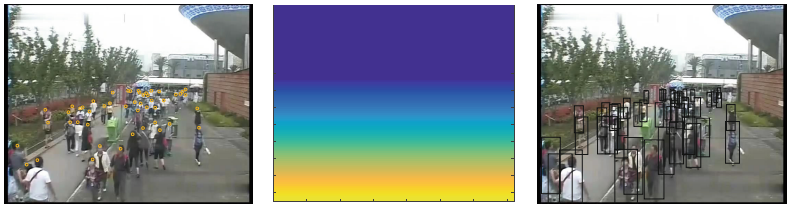
3 Experiments

In the experiments, we used intel $i5$-2500 CPU, 16 GB RAM, and NVIDIA GTX970 GPU. The implementation of the proposed system was based on **matConvNet** library [19] and the experiments were performed using **MATLAB 2015B**.

3.1 Implementation Details

The image size was set to 576×720 in both training and testing procedures and the convolutional feature map was designed to have ($18 \times 23 \times 512$) dimensions. All the convolutional layers from VGG-16 [17] were initialized by the pre-trained model for ImageNet classification [15]. The fully-connected layers in both PPN and DASN were initialized with a zero-mean Gaussian distribution with standard deviation 0.01.

The proposed network was trained on *WorldExpo* dataset containing 3374 training images. The parameter settings are illustrated in Table 1. The learning rate η to 10^{-6} for the first 20 epochs of the training data whereas to 10^{-7} for next 40 epochs. The training parameters for our network were determinded by



(a) (b) (c)

Fig. 4. An example of generating bounding boxes using annotated head positions and perspective map of *WorldExpo'10 dataset* [20]. **(a)** the annotated head positions of pedestrians. **(b)** the given perspective map. **(c)** the estimated bounding boxes.

referring to the settings of Faster R-CNN [14]. The weight parameters of the multi-task loss function in (1) were introduced to adjust the relative importance of each loss term. Total $18 \times 23 \times 24 = 9936$ predicted proposals were extracted per image but the number of true pedestrians were hundreds at most. Therefore, the negative samples became dominant and this would lead to biased training results. To avoid the bias, we randomply sampled equal number of positive proposals and negative proposals (128 each) as for the training samples.

3.2 Experiments on WorldExpo'10 Dataset

We developed two-step training scheme to train the proposed network. We first trained PPN and the convolutional networks by back-propagation and stochastic gradient descent (SGD) with the ground truth bboxes and density map. In this step, the PPN is trained to produce initial proposals and crowd density map. After the training the PPN, then we trained the DASN using the ground truth bboxes to suggest the final proposal scores. In this step, the PPN was fixed and only the two fully-connected layers of the DASN were trained by back-propagation and SGD.

Table 1. Parameter Settings

Symbol	Definition	Values
λ_c	Weight of the confidence score loss	1
λ_b	Weight of the regression loss	1
λ_d	Weight of the density estimation loss	$\frac{1}{128}$
N_p	Number of samples (proposal network)	256
N_s	Number of samples (selection network)	256
η	Learning rate	10^{-6} to 10^{-7}
γ	Weight decay	0.0005
μ	Momentum	0.9

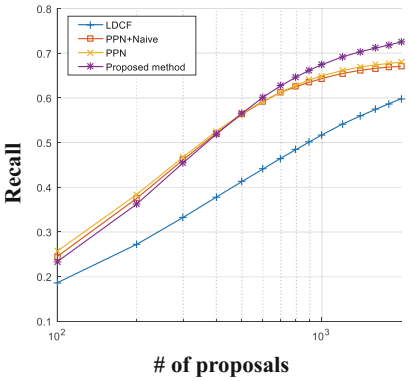


Fig. 5. Comparison of different methods in pedestrian proposal on *WorldExpo'10* dataset.

The Ground Truths. The *WorldExpo'10 dataset* [20] is a crowd counting dataset including 3875 images from 108 scenes, which provides the annotated head positions of people and the perspective map of scenes. Although this dataset is an excellent large-scale dataset for crowd counting, it has no annotated bounding boxes which is essential for the purpose of pedestrian detection. Therefore, we estimate bounding boxes using the given head position and perspective maps of the dataset and use the boxes as the ground truth. As shown in Fig. 4, the locations and sizes of the crowd people are easily estimated by predicting the size of pedestrian using the perspective map. We assumed the aspect ratio of pedestrian should be consistent and fixed by 3 (width/height). In addition, since the proposed method utilizes the crowded density map, the ground truths of the crowd density should be constructed. The density map of an image has 18×23 cells and each cell counts the number of ground truth bboxes of the image whose center is inside the cell.

Processing Times. One strong point of our method is the speed much faster than that of LDCF [11] or EdgeBoxes [21]. Since the proposed network forms a feed-forward network, the forward pass for a test image (576×720) takes about 40 ms on a GPU (25 fps).

Evaluation. We used 3376 images for training the proposed network and 599 images for testing in the same setting of [20]. The trained DAPPN was also used for the crowd-scene dataset *PETS2009* [1]. To evaluate the effectiveness of our algorithm, we compared various methods such as LDCF, PPN, PPN+Naive, and DAPPN (the proposed method). We trained the LDCF detector using training images of *WorldExpo'10*. Since the performance of LDCF is affected by the size of test images, we empirically resized the test images two times to achieve the best performance of LDCF. In order to validate the pedestrian proposal networks (PPN) in Sect. 2.1, the initial proposals from PPN were evaluated. In addition,

we developed a naive approach (PPN+Naive) which re-weights the confidence scores of PPN by multiplying their density values.

The quantitative results are shown in Fig. 5, where the recall performance (IOU > 0.5) is depicted versus the number of proposals per image. As shown in the Fig. 5, PPN, PPN+Naive, and the proposed method outperformed LDCF detector. PPN and PPN+Naive show almost the same results even though PPN+Naive approach re-weights the confidence scores of PPN using the crowd density. The reason is that since the ground truths of crowd density are provided into PPN in training phase, the trained PPN contains enough crowd density information. Figure 7 shows the qualitative comparisons of LDCF and the proposed method. The black boxes mean the missed pedestrians and the red boxes represent the correctly detected pedestrians. The top-300 scored pedestrian proposals are illustrated in Fig. 7. The proposed method shows the better proposal performance than LDCF in the crowded region. On the contrary, LDCF has an advantage to detect the isolated pedestrians in non-crowded region. Since the proposed method depend on the crowd density, the confidence scores of the proposals in low density region are tend to be low.

3.3 Experiments on PETS2009 Dataset

The purpose of *PETS2009* [1] dataset is to provide crowded scenes and people annotations to analyze the behaviors of the occluded people. We performed experiments in S2L2 and S2L3 scenes with 656 images. In these scenes, more than 20 people are walking around with heavy occlusions. The test images were resized to 576×720 to fit the proposed network input size. The qualitative and quantitative results are shown in Figs. 6 and 8 respectively. The test settings for LDCF detector and the proposed method were the same as Sect. 3.2. The recall performance of the proposed method outperformed PPN and LDCF [11]. As shown in Fig. 8, our method successfully localized the pedestrians in the situation of severe occlusions while LDCF failed to find the occluded people.

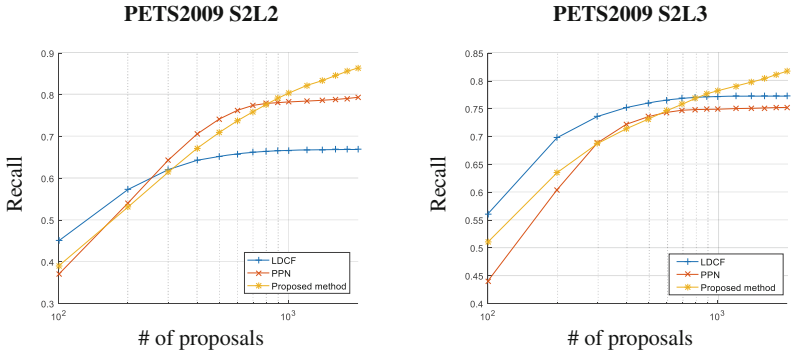


Fig. 6. Comparison of different methods in pedestrian proposal on *PETS2009* dataset.

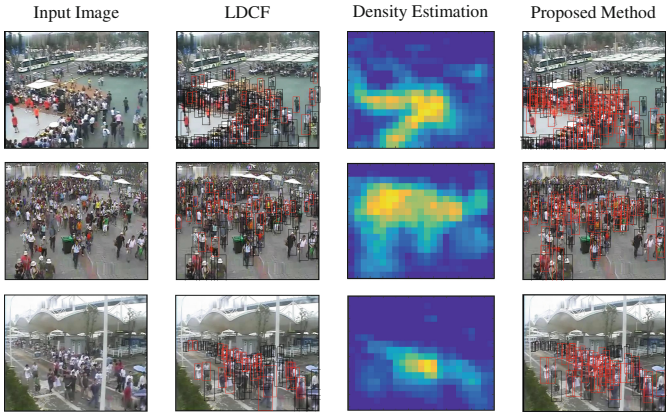


Fig. 7. Examples of the qualitative results on *WorldExpo'10* dataset. The black colored bounding boxes denote the missed ground truths and the red boxes mean the true positives. (Color figure online)

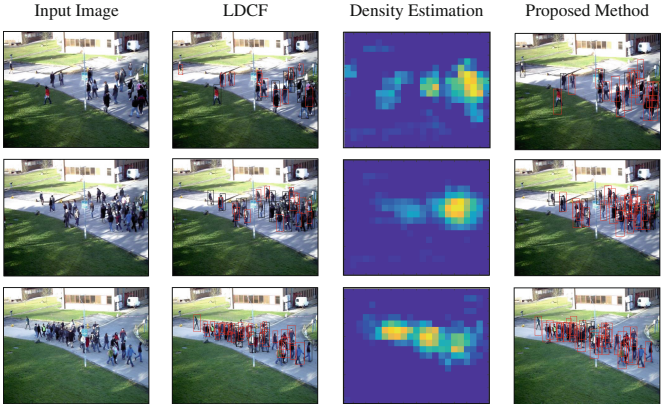


Fig. 8. Examples of the qualitative results on *PETS2009* dataset.

4 Conclusion

In this paper, we proposed a density-aware pedestrian proposal network to solve the occlusion problems in crowded scenes. Since the traditional pedestrian detectors [6,11] or object proposal methods [14,21] try to find pedestrians in a crowded scene without considering the occlusions, performance degradation often occurs. The proposed method utilizes a crowd density map in order to generate pedestrian proposals robust against severe occlusions. We have designed the proposal network to produce initial proposals and the selection network to determine the final proposals using crowd density. We have evaluated the proposed method on *WorldExpo10* [20] and *PETS2009* [1] datasets. The experiments verified the

effectiveness of the proposed method by comparing it with the conventional method LDCF [11].

References

1. http://www.cvg.reading.ac.uk/PETS2009/a.html (2009)
2. Arbeláez, P., Pont-Tuset, J., Barron, J., Marques, F., Malik, J.: Multiscale combinatorial grouping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 328–335 (2014)
3. Carreira, J., Sminchisescu, C.: CPMC: automatic object segmentation using constrained parametric min-cuts. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1312–1328 (2012)
4. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE Trans. Pattern Anal. Mach. Intell. **36**(8), 1532–1545 (2014)
5. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes challenge 2012 (VOC2012) results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
6. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008), pp. 1–8. IEEE (2008)
7. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
8. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
10. Li, J., Liang, X., Shen, S., Xu, T., Yan, S.: Scale-aware fast R-CNN for pedestrian detection. arXiv preprint (2015). arXiv:1510.08160
11. Nam, W., Dollár, P., Han, J.H.: Local decorrelation for improved pedestrian detection. In: Advances in Neural Information Processing Systems, pp. 424–432 (2014)
12. Ouyang, W., Wang, X.: Single-pedestrian detection aided by multi-pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3198–3205 (2013)
13. Ouyang, W., Zeng, X., Wang, X.: Modeling mutual visibility relationship in pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3222–3229 (2013)
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
15. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015)
16. Van de Sande, K.E., Uijlings, J.R., Gevers, T., Smeulders, A.W.: Segmentation as selective search for object recognition. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1879–1886. IEEE (2011)

17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint (2014). arXiv:1409.1556
18. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1904–1912 (2015)
19. Vedaldi, A., Lenc, K.: Matconvnet - convolutional neural networks for matlab (2015)
20. Zhang, C., Li, H., Wang, X., Yang, X.: Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 833–841 (2015)
21. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 391–405. Springer, Heidelberg (2014)