Journal of Theoretical and Applied Information Technology
15th May 2020. Vol.98. No 09

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REAL-TIME APPROACH FOR ACCURATE PEDESTRIAN LOCALIZATION IN CROWDED SCENES

HOANH NGUYEN
Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam
E-mail: nguyenhoanh@iuh.edu.vn

ABSTRACT

Detection of pedestrians in a crowded scene is a difficult problem since pedestrians usually gather and occlude each other. In addition, due to the complexity of crowded scenes, current deep learning-based approaches for pedestrian detection still require high computational cost. This paper addresses above problems by introducing a deep learning-based approach for fast and accurate pedestrian detection in a crowded scene. To reduce computational cost and increase inference speed, a reduced ShuffleNet network based on ShuffleNet architecture is first adopted as the base network to generate the base convolution layers. ShuffleNet architecture is built on ShuffleNet units and Strided ShuffleNet units, which include pointwise group convolution layers and channel shuffle operations to greatly reduce computation cost while maintaining detection accuracy. To solve the issue of highly overlapped pedestrian in crowded scenes, an improved non-maximum suppression algorithm is developed based on density score map generated by density prediction sub-network. The improved non-maximum suppression algorithm proposes a dynamic suppression strategy, where the threshold value for suppression rises as pedestrian instances gather and occlude each other and decays when pedestrian instances appear separately. Experimental results on CityPersons dataset and CrowdHuman dataset show the effectiveness of the proposed approach on pedestrian detection in crowded scenes.

Keywords: *Pedestrian Detection, Deep Learning, Non-Maximum Suppression, ShuffleNet Network, Channel Shuffle*

1. INTRODUCTION

Vision-based pedestrian detection is a key problem in many intelligent transport systems. In video surveillance systems, pedestrian detection can provide fundamental information for people counting, event recognition, and crowd monitoring. In intelligent transportation, pedestrian detection is an essential part for the semantic understanding of the environment. Pedestrian detection can be seen as an aspect of generic object detection. Early generic object detection approaches rely on the sliding window paradigm based on the hand-crafted features and classifiers to locate objects. Dollár et al. [20] proved that using features from multiple channels can significantly improve the detection performance. In [21], the authors proposed to combine a margin-sensitive approach for data-mining hard negative examples with a formalism. Viola et al. [22] proposed a new image representation which allows the features to be computed very quickly. In recent years, with the advent of deep convolutional neural network (CNN), a new generation of more effective object detection methods based on CNN significantly improve the detection performance. Deep CNN-based generic object detection approaches can be roughly classified into two categories: one-stage and two-stage approaches. One-stage object detection approaches, such as YOLO [24] and SSD [23], operate in a sliding window manner that makes prediction for densely sampled locations in the input image. Two-stage object detection approaches, such as Faster R-CNN [11] and R-FCN [25], first generate a set of region proposals and then perform a second stage prediction to classify each proposal and refine the bounding box of proposals. Two-stage approaches achieve better detection accuracy than one-stage approaches, but are significantly slower [26]. Two-stage object detection approaches normally consist of a region proposal network (RPN) that hypothesizes candidate object locations and a detection network that refines

Journal of Theoretical and Applied Information Technology
15th May 2020. Vol.98. No 09

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

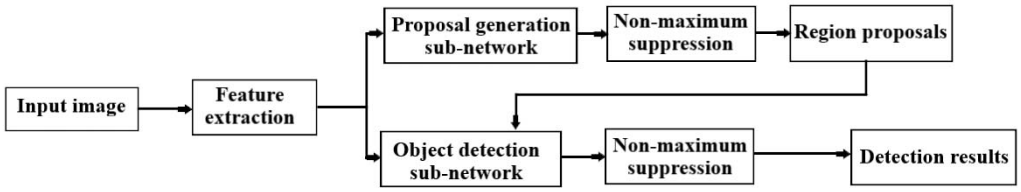

Figure 1: The Typical Blocks of Two-Stage Object Detectors.

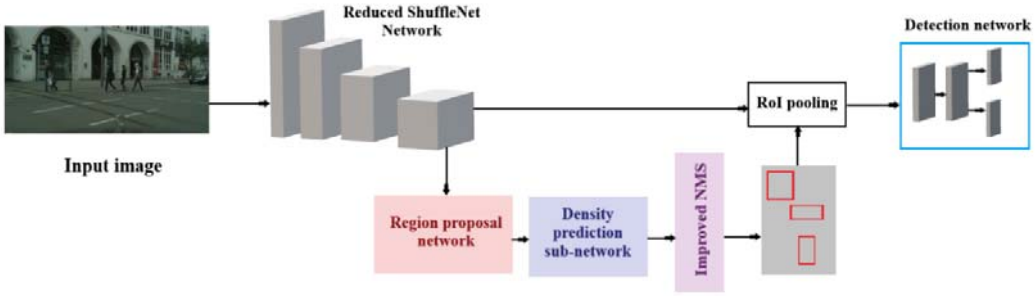

Figure 2: The Overall Structure of The Proposed Method.

region proposals. The RPN and detection network share the same feature extractor network. Figure 1 illustrates the typical blocks of two-stage object detectors. As in [11], most computational cost happens in feature extraction stage.

In terms of pedestrian detection, driven by the success of R-CNN [27], a series of pedestrian detection frameworks are proposed in the two-stage approach. The author in [28] proposed to use deconvolutional modules to bring additional context information which is more effective to detect small-scale pedestrians. Sermanet et al. [29] presented an unsupervised method using the convolutional sparse coding to pre-train CNN for pedestrian detection. In [30], a complexity-aware cascaded detector was proposed for an optimal trade-off between accuracy and speed. Angelova et al. [31] combined the ideas of fast cascade and a deep network to detect pedestrian. Yang et al. [32] used scale-dependent pooling and layer-wise cascaded rejection classifiers to detect objects efficiently. Zhang et al. [33] presented an effective pipeline for pedestrian detection via using RPN followed by boosted forests. Li et al. [34] use multiple built-in sub-networks to adaptively detect pedestrians across scales. Although numerous pedestrian detection methods are presented in literature, how to robustly detect each individual pedestrian in crowded scenes is still one of the most critical issues for pedestrian detectors.

2. METHOD

Figure 2 illustrates the overall architecture of the proposed method. To enhance the performance of pedestrian detection in crowded scenes in both detection accuracy and inference speed, a reduced ShuffleNet network based on ShuffleNet architecture is first adopted to generate base convolution layers. ShuffleNet architecture is built on ShuffleNet units and Strided ShuffleNet units, which include pointwise group convolution layers and channel shuffle operations to greatly reduce computation cost while maintaining detection accuracy. To solve the issue of highly overlapped pedestrian in crowded scenes, an improved non-maximum suppression algorithm is developed based on density score map generated by density prediction sub-network. The improved non-maximum suppression algorithm proposes a dynamic suppression strategy, where the threshold value for suppression rises as pedestrian instances gather and occlude each other and decays when pedestrian instances appear separately. Details of the proposed model is explained in the following sections.

2.1 ShuffleNet Architecture

ShuffleNet [9] is an extremely computation-efficient deep convolutional neural network (CNN) architecture. ShuffleNet adopts two operations, including pointwise group convolution and channel shuffle, to greatly reduce computation cost while

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

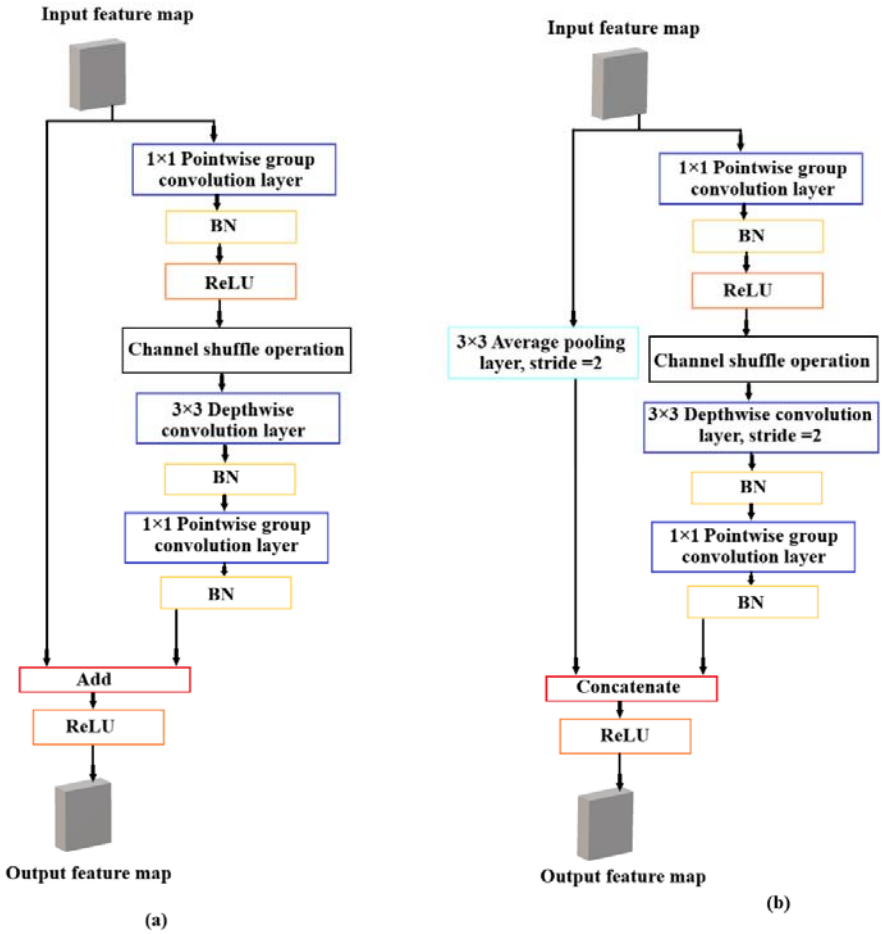

Figure 3: The Structure of the ShuffleNet Unit (a) and the Strided ShuffleNet Unit (b).

Table 1: The Architecture of the ShuffleNet Used in This Paper. The Size of Input Image Is 224×224.

Layer	Type	#Repeat	Kernel size	Output size
0	Standard convolution, stride 2	1	3×3	112×112
	Max pooling	-	3×3	56×56
1	Strided ShuffleNet unit	1	-	28×28
	ShuffleNet unit	3	-	28×28
2	Strided ShuffleNet unit	1	-	14×14
	ShuffleNet unit	7	-	14×14
3	Strided ShuffleNet unit	1	-	7×7
	ShuffleNet unit	3	-	7×7

maintaining accuracy. ShuffleNet architecture is built on ShuffleNet units and Strided ShuffleNet units. Figure 3 (a) illustrates the structure of the ShuffleNet unit. It is based on the principle of residual block [1]. As shown in Figure 3 (a), in the residual branch of ShuffleNet unit, a 1×1 pointwise group convolution layer followed by a channel shuffle operation is first added. Then, a computational economical 3×3 depthwise convolution layer [2] is applied on the bottleneck feature map. Finally, a second pointwise group convolution layer is adopted after depthwise convolution layer to recover the channel dimension to match the shortcut path. Batch normalization (BN) [3] is used after each convolution layer, and ReLU layer is used after the first pointwise group

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

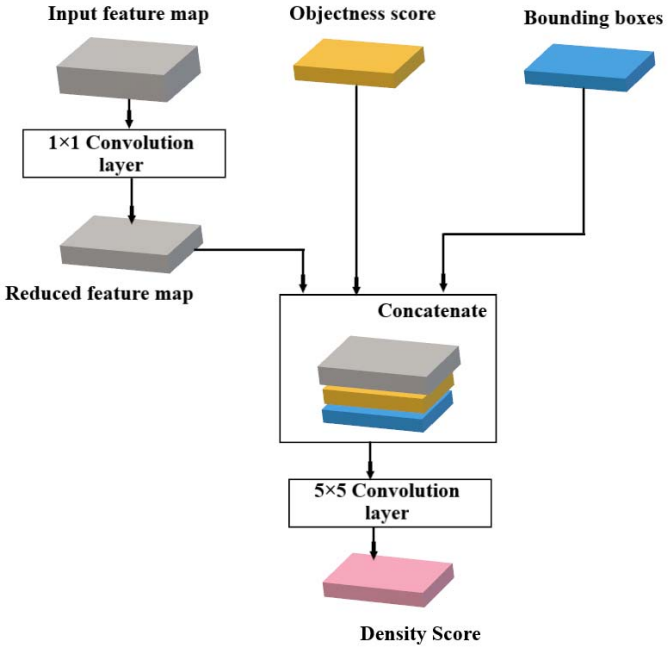

Figure 4: The Structure of The Proposed Density Sub-Network.

convolution layer. For the Strided ShuffleNet unit, a 3×3 depthwise convolution layer with stride =2 is used instead of standard depthwise convolution layer in ShuffleNet unit. In addition, a 3×3 average pooling layer is added on the shortcut path of the Strided ShuffleNet unit, and the element-wise addition in the ShuffleNet unit is replaced by channel concatenation in the Strided ShuffleNet unit, which makes it easy to enlarge channel dimension with little extra computation cost. Figure 3(b) illustrates the structure of the Strided ShuffleNet unit. The channel shuffle operation used in the ShuffleNet units and the Strided ShuffleNet units first reshapes the output channel dimension, and then transposes and flattens it back as the input of next layer. The channel shuffle operation makes it possible to build more powerful structures with multiple group convolutional layers.

Based on the ShuffleNet units and the Strided ShuffleNet units, the architecture of the ShuffleNet used in this paper is shown as in Table 1. As shown in Table 1, ShuffleNet consists of a stack of ShuffleNet units and Strided ShuffleNet units grouped into three layers (layers 2 to layers 4). At the first layer, a standard 3×3 convolution layer followed by max polling layer is applied on the input image. With pointwise group convolution and channel shuffle operation, all components in ShuffleNet can be computed efficiently. Compared with ResNet [1] and ResNeXt [4], the ShuffleNet structure has less complexity under the same settings.

2.2 The Density Sub-Network for Density Prediction

Occlusion issues are the main problem of detecting pedestrian in a crowd scene. Recently, Non-Maximum suppression (NMS) algorithm, soft-NMS algorithm [1] and learning NMS algorithm [14] are usually adopted to solve the occlusion problem in object detection. However, the highly overlapped pedestrian in crowd scenes leads to missing highly overlapped objects. To solve this problem, this paper adopts an auxiliary and learnable sub-network to predict the NMS threshold for solving the occlusion issues in crowd scenes. With the proposed density sub-network, a dynamic suppression strategy, where the NMS threshold rises as pedestrian instances gather or occlude each other and decays when pedestrian instances appear separately, is introduced in the next sub-section. Figure 4 illustrates the structure of the proposed density sub-network. The density sub-network includes three convolutional layers to predict the density of each proposal. The density sub-network is constructed behind the RPN. As shown, a 1×1 convolutional layer is first applied to reduce the dimension of the convolutional feature maps. The reduced feature maps then are concatenated with the objectness score and the bounding boxes predicted by the RPN. The output of the concatenation

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

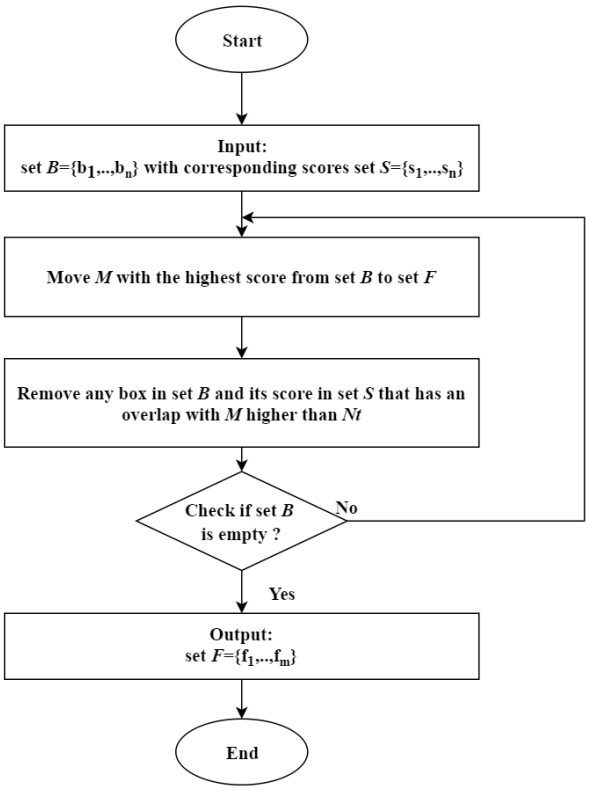

Figure 5: The Flowchart of Traditional NMS Algorithm.

operation is fed into the density sub-network. In addition, a 5×5 kernel is used at the final convolutional layer of the density sub-network to take the surrounding information into account, thus enhancing the density prediction.

2.3 Improved NMS Algorithm for Solving the Occlusion Issues

2.3.1 Traditional NMS Algorithm

Traditional NMS algorithm was first demonstrated in [5] to surpass other approaches for human detection. Since then, it has been a standard component in object detection approaches and widely used in one-stage and two-stage detectors such as SSD, Faster R-CNN. Figure 5 shows the flowchart of traditional NMS algorithm. As shown in Figure 5, traditional NMS algorithm starts with a set of detection boxes $B = \{b_1, b_2, \dots, b_n\}$ with corresponding scores $S = \{s_1, s_2, \dots, s_n\}$, NMS firstly selects the one M with the maximum score and moves it from set B to the set of final detections $F = \{f_1, f_2, \dots, f_n\}$. It then removes any box in set B and its score in set S that has an overlap with M higher than a manually set threshold N_t. This process is repeated for the remaining B set.

The detection performance of traditional NMS-based approaches is usually based on the value of threshold N_t. If the value of threshold N_t is low, the miss-rate, which shows the ability of the detector for balancing recall and precision, may increase, especially in crowd scenes. Figure 6 shows an example of pedestrian detection results in this case. As shown in Figure 6, there may be many pairs of crowded pedestrians which have higher overlaps than the suppressing threshold N_t. Within these pairs, if the bounding box with the maximum score M is selected, all the surrounding detected boxes that have overlaps greater than threshold N_t are suppressed, including the nearby detected pedestrians that actually locate the other ground truth pedestrians. In this case, true positives may be removed after the NMS processing with a low threshold N_t, thus increasing the miss-rate. On the other hand, if the value of threshold N_t is high, the false positives may increase as many surrounding detected boxes that are overlapped often have correlated scores. Although more highly overlapped true positives can be kept with higher value of threshold N_t, the increase of false positives may be more serious because the number of pedestrians is typically smaller than the number of proposals

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

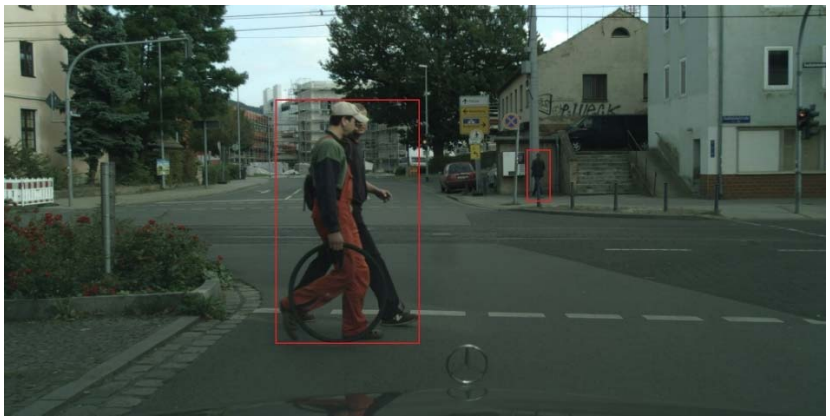

Figure 6: Example of Pedestrian Detection Results with Low NMS Threshold.

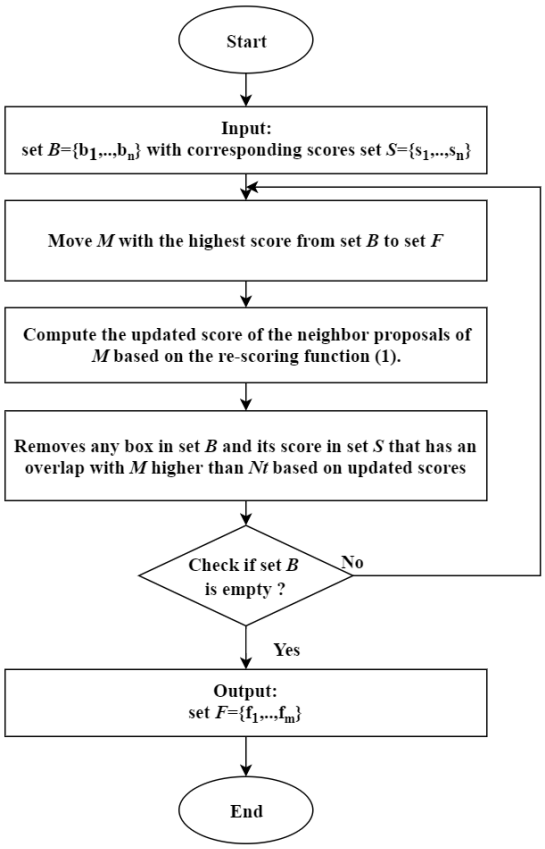

Figure 7: The Flowchart of Soft-NMS Algorithm.

generated by a detector. Therefore, using a high NMS threshold is not a good choice either.

2.3.2 Soft-NMS Algorithm

To solve the problem with traditional NMS algorithm on occlusion issues, Soft-NMS algorithm [1] defined the re-scoring function as follow:

$$S_i' = \begin{cases} s_i & if\ (IoU(M, b_i)) < N_t \\ s_i.(1 - IoU(M, b_i)) & if\ (IoU(M, b_i)) \geq N_t \end{cases} \quad (1)$$

where S_i' denotes the updated objectiveness score of b_i. Figure 7 shows the flowchart of Soft-NMS algorithm based on the re-scoring function. With Soft-NMS, the neighbor proposals of the bounding box with the maximum score M are not completely suppressed. Instead they are suppressed based on updated objectiveness scores of the neighbor proposals, which are computed according to the overlap level of the neighbor proposals and the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

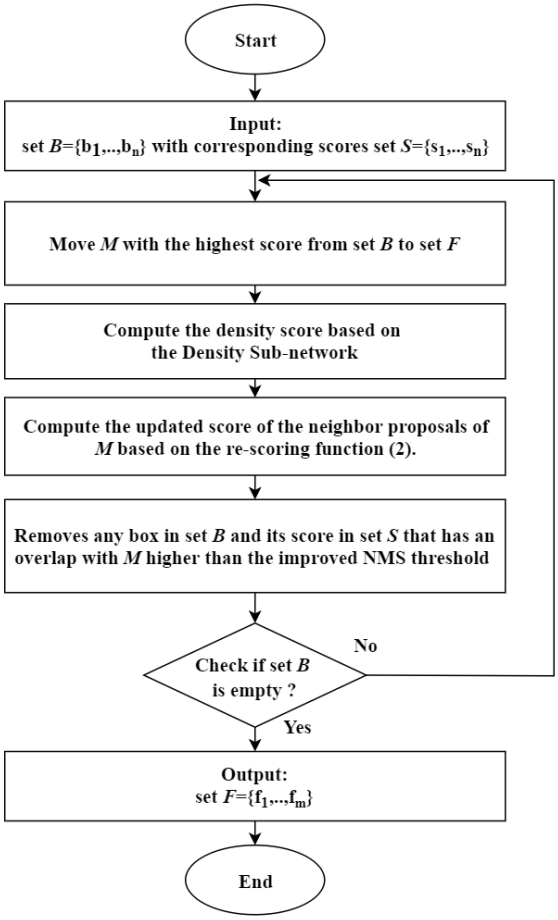

Figure 8: The Flowchart of The Improved NMS Algorithm.

current proposal. In general, Soft-NMS is a variant of traditional NMS that decays the score of neighboring detections instead of totally removing it. However, Soft-NMS still assigns a greater penalty to the highly overlapped boxes, which approximately equals to that in traditional NMS.

2.3.3 Improved NMS Algorithm

Based on the above analysis with both traditional NMS and Soft-NMS, increasing the threshold value in NMS-based algorithm with highly overlapped pedestrians in a crowded region seems can preserve neighboring detections. In addition, in the sparse region, highly overlapped pedestrians should be removed, as they are more likely to be false positives. To conducts an improved suppressing algorithm based on above conclusions, this paper first defines a re-scoring function as follow:

$$S_i' = \begin{cases} s_i & if\ (IoU(M, b_i)) < N_\propto \\ s_i.\left(1 - IoU(M, b_i)\right) & if\ (IoU(M, b_i)) \geq N_\propto \end{cases} \quad (2)$$

where $N_\propto$ denotes the improved NMS threshold for M; $N_\propto$ is defined as follow equation:

$$N_\propto = \max(N_t, \max(IoU(b_i, b_j))) \quad (3)$$

where $\max(IoU(b_i, b_j))$ is the density score of the object i; the density score is the max bounding box IoU of current object with other objects in the ground truth set. The density score of objects indicates the level of crowd occlusion.

Figure 8 illustrates the flowchart of the improved NMS algorithm based on the density score. If $IoU(M, b_i) < N_t$ (the neighboring bounding boxes are far away from M), they are kept the same as the traditional NMS does. If $\max(IoU(b_i, b_j)) > N_t$ (M locates in the crowded region), the neighboring bounding boxes are preserved based on the density score. If $\max(IoU(b_i, b_j)) \leq N_t$ (M locates in the sparse region), the NMS threshold $N_\propto$ equals to N_t. Hence, the suppressing step is equivalent to the traditional NMS, where very close bounding boxes are

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Comparisons of Detection Results with Previous State-of-The-Art Methods on CityPersons Dataset.

Method	Backbone Network	MR^{-2} (%)				
		Reasonable	Heavy	Partial	Bare	Processing Time (s)
Proposed method	ShuffleNet	12.9	54.2	14.6	9.2	0.11
OR-CNN [15]	VGG-16	12.8	55.7	15.3	6.7	-
Repulsion Loss [14]	ResNet-50	13.2	56.9	16.8	7.6	-
TLL [17]	ResNet-50	15.5	53.6	17.2	10.0	-
ALFNet [16]	ResNet-50	12.0	51.9	11.4	8.4	0.27

Table 3: Detection Results of The Proposed Experiments on CrowdHuman Dataset.

Network	MR^{-2} (%)
FPN with Faster R-CNN baseline [19]	50.42
FPN + ShuffleNet	49.18
FPN + ShuffleNet + Soft-NMS	47.98
FPN + ShuffleNet + improved NMS	47.62

suppressed as false positives. Thus, the improved NMS algorithm used in this paper applies a dynamic suppression strategy, where the threshold rises as instances gather and occlude each other and decays when instances appear separately.

3. EXPERIMENTS

The proposed approach is implemented in Pytorch deep-learning framework with Python interface. The CPU used in all experiments is Intel Core i7-8700, the main memory is 12GB DDR4 RAM, and the GPU is NVIDIA GeForce GTX 1080. Two widely used datasets are adopts to evaluate the proposed method, including CityPersons dataset [7] and CrowdHuman dataset [8]. The proposed network follows the Faster R-CNN framework [11] and uses ShuffleNet [9], pre-trained on the MS COCO dataset [10], as the backbone network. All the parameters in the newly added convolutional layers are randomly initialized by the "xavier" method [12]. The proposed network is optimized by using the Stochastic Gradient Descent algorithm with 0.9 momentum and 0.0005 weight decay, the mini-batch is set at one image per GPU. For the CityPersons dataset, this paper sets the learning rate to 10−3 for the first 20k iterations, and decay it to 10−4 for another 30k iterations. For the CrowdHuman dataset, the initial learning rate is set at 10−3 for 10k iterations, and decrease by a factor of 10 after the first 20k iterations.

3.1 Experiments on CityPersons Dataset

The CityPersons dataset is built upon the semantic segmentation dataset Cityscapes [13] to provide a new dataset for pedestrian detection with heavy occlusion. The dataset was recorded across 18 different cities in Germany with three different seasons and various weather conditions. The dataset includes 5,000 images with 2,975 images for training, 500 images for validation, and 1,525 images for testing. In this paper, the proposed network is trained on the training set and evaluated on the validation set. Following the evaluation metrics in CityPersons dataset [7], the log miss rate averaged over the false positive per image (FPPI) range of $[10^{-2}, 10^{0}]$ (denoted as MR^{-2}, lower score indicates better performance) is used to evaluate the detection performance. MR^{-2} is computed by averaging the Miss Rate at 9 FPPI rates over the range of $[10^{-2}, 10^{0}]$ in log-space. In addition, this paper follows the strategy in [14] and [15] to divide the reasonable subset in the validation set into the Partial subset (10% < occlusion < 35 %), Bare subset (occlusion ≤ 10%) subsets and Heavy subset (occlusion ≥ 35%).

To evaluate the performance of the proposed approach, this paper conducts experiments on CityPersons dataset and compares the detection performance with recent pedestrian detection approaches, including OR-CNN [15], Repulsion Loss [14], ALFNet [16], and TLL [17]. OR-CNN designed a new aggregation loss to enforce proposals to be close and locate compactly to the corresponding objects, and a new part occlusion-aware region of interest pooling unit to replace the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 9: Detection Results on CityPersons Dataset.

RoI pooling layer in order to integrate the prior structure information of human body with visibility prediction into the network to handle occlusion. Repulsion Loss proposed a novel bounding box regression loss specifically designed for crowd scenes, which prevents the proposal from shifting to surrounding objects thus leading to more crowd-robust localization. ALFNet proposed a novel architecture which stacks a series of predictors to directly evolve the default anchor boxes of SSD step by step into improving detection results. TLL proposed a novel method integrated with somatic topological line localization and temporal feature aggregation for detecting multi-scale pedestrians, which works particularly well with small-scale pedestrians that are relatively far from the camera.

Table 2 shows the comparisons of detection results with previous state-of-the-art methods on CityPersons dataset. It can be observed that the proposed method achieves comparable performance compared with the state-of-the-art methods on all subsets. More specific, the MR^{-2} of the proposed method is 12.9% with Reasonable subset, 54.2% with Heavy subset, 14.6% with Partial subset, and 9.2% with Bare subset. For the processing speed, it can be seen that the proposed method can achieve significant speed-up. The proposed method takes only 0.11 second per image, which meets the requirement of real-time detection in intelligent transport systems. The above results show that the proposed approach is effective in both detection accuracy and inference speed, and has a good

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 10: Visual Results of The Proposed Method on CrowdHuman Dataset.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

potential for processing detectors in crowd scenes. Figure 9 shows detection results on CityPersons dataset. It can be observed that the proposed method is superior on detecting pedestrians of various scales and occlusion levels.

3.2 Experiments on CrowdHuman Dataset

To evaluate the effectiveness of each proposed module, this paper conducts experiments on CrowdHuman dataset. CrowdHuman dataset [8] has been introduced recently to specifically target to the crowd issue in the human detection task. The dataset includes 24,370 images with 15,000 images for training, 4,370 images for validation and 5,000 images for testing. There are 340,000 pedestrian annotations and 99,000 ignore region annotations in the training set. In addition, the dataset contains approximately 22.6 pedestrians in average per image as well as 2.4 pairwise crowd instances. This paper follows the evaluation metrics used in CrowdHuman [8], denoted as MR^{-2}. All the experiments are trained in the CrowdHuman training set and evaluated in the validation set. All results are compared to original FPN [19] to show the effectiveness of each proposed module. Three experiments are conducted on CrowdHuman dataset. First, the VGG-16 [18] network in FPN is replaced by ShuffleNet architecture proposed in this paper. Second, the NMS algorithm in FPN is replaced by Soft-NMS algorithm. Finally, the improved NMS algorithm with the density prediction sub-network proposed in this paper is adopted to replace NMS algorithm.

Table 3 shows the detection results of the proposed experiments and FPN. It can be observed from Table 3 that using the ShuffleNet network as the base network, the proposed approach achieves 49.18% MR−2 on the validation set, which is better than the reported result of FPN with Faster R-CNN baseline [19] (50.42%). The soft-NMS algorithm slightly reduces the MR−2 by 1.2% for FPN + ShuffleNet framework. Combining improved NMS algorithm with FPN + ShuffleNet framework, the MR−2 score drops to 47.62% with a 0.36% reduction compared with soft-NMS. These results indicate that adaptive-NMS keeps more true positives, and it is a more effective postprocessing algorithm for detecting pedestrians in crowded scenarios. Figure 10 shows some visual results of the proposed method on CrowdHuman dataset. It can be seen that the proposed network keeps more crowded true positives and still removes false positives in the sparse region at the same time.

4. CONCLUSIONS

This paper presents a deep learning-based framework for pedestrian detection in a crowded scene. In the proposed framework, a reduced ShuffleNet network is used to generate base convolution layers, which enhance the performance of pedestrian detection in crowded scenes in both detection accuracy and inference speed. Meanwhile, to solve the issue of highly overlapped pedestrian in crowded scenes, an improved non-maximum suppression algorithm is developed based on density score map generated by density prediction sub-network. The improved non-maximum suppression algorithm proposes a dynamic suppression strategy, where the threshold value for suppression rises as pedestrian instances gather and occlude each other and decays when pedestrian instances appear separately. Extensive experiments on CityPersons dataset and CrowdHuman dataset with popular pedestrian detection frameworks demonstrated the effectiveness and robustness of the proposed method. More specific, the proposed approach achieved comparable detection accuracy while being faster. In the future, this paper plans to extend the proposed method to detect other kinds of objects such as car, bicycle, tricycle.

REFERENCES:

[1] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[2] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251-1258. 2017.

[3] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167* (2015).

[4] Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated residual transformations for deep neural networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492-1500. 2017.

[5] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." In *2005 IEEE computer society conference on*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

computer vision and pattern recognition (CVPR'05), vol. 1, pp. 886-893. IEEE, 2005.

[6] Bodla, Navaneeth, Bharat Singh, Rama Chellappa, and Larry S. Davis. "Soft-NMS--improving object detection with one line of code." In *Proceedings of the IEEE international conference on computer vision*, pp. 5561-5569. 2017.

[7] Zhang, Shanshan, Rodrigo Benenson, and Bernt Schiele. "Citypersons: A diverse dataset for pedestrian detection." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213-3221. 2017.

[8] Shao, Shuai, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. "Crowdhuman: A benchmark for detecting human in a crowd." *arXiv preprint arXiv:1805.00123* (2018).

[9] Zhang, Xiangyu, Xinyu Zhou, Mengxiao Lin, and Jian Sun. "Shufflenet: An extremely efficient convolutional neural network for mobile devices." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848-6856. 2018.

[10] Lin, Tsung-Yi, 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*, pp. 740-755. Springer, Cham, 2014.

[11] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." In *Advances in neural information processing systems*, pp. 91-99. 2015.

[12] Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249-256. 2010.

[13] Cordts, Marius, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. "The cityscapes dataset for semantic urban scene understanding." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213-3223. 2016.

[14] Wang, Xinlong, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. "Repulsion loss: Detecting pedestrians in a crowd." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7774-7783. 2018.

[15] Zhang, Shifeng, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. "Occlusion-aware R-CNN: detecting pedestrians in a crowd." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 637-653. 2018.

[16] Liu, Wei, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 618-634. 2018.

[17] Song, Tao, Leiyu Sun, Di Xie, Haiming Sun, and Shiliang Pu. "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 536-551. 2018.

[18] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[19] Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature pyramid networks for object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117-2125. 2017.

[20] Dollár, Piotr, Ron Appel, Serge Belongie, and Pietro Perona. "Fast feature pyramids for object detection." *IEEE transactions on pattern analysis and machine intelligence* 36, no. 8 (2014): 1532-1545.

[21] Felzenszwalb, Pedro F., Ross B. Girshick, David McAllester, and Deva Ramanan. "Object detection with discriminatively trained part-based models." *IEEE transactions on pattern analysis and machine intelligence* 32, no. 9 (2009): 1627-1645.

[22] Viola, Paul, and Michael J. Jones. "Robust real-time face detection." *International journal of computer vision* 57, no. 2 (2004): 137-154.

[23] Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. "Ssd: Single shot multibox detector." In *European conference on computer vision*, pp. 21-37. Springer, Cham, 2016.

[24] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779-788. 2016.

[25] Dai, Jifeng, Yi Li, Kaiming He, and Jian Sun. "R-fcn: Object detection via region-based fully

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

convolutional networks." In *Advances in neural information processing systems*, pp. 379-387. 2016.

[26] Huang, Jonathan, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer et al. "Speed/accuracy trade-offs for modern convolutional object detectors." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7310-7311. 2017.

[27] Girshick, Ross, 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*, pp. 580-587. 2014.

[28] NGUYEN, HOANH. "AN EFFICIENT DEEP LEARNING FRAMEWORK FOR PEDESTRIAN DETECTION." *Journal of Theoretical and Applied Information Technology* 97, no. 21 (2019).

[29] Sermanet, Pierre, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun. "Pedestrian detection with unsupervised multi-stage feature learning." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3626-3633. 2013.

[30] Cai, Zhaowei, Mohammad Saberian, and Nuno Vasconcelos. "Learning complexity-aware cascades for deep pedestrian detection." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3361-3369. 2015.

[31] Angelova, Anelia, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, and Dave Ferguson. "Real-time pedestrian detection with deep network cascades." (2015).

[32] Yang, Fan, Wongun Choi, and Yuanqing Lin. "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2129-2137. 2016.

[33] Zhang, Liliang, Liang Lin, Xiaodan Liang, and Kaiming He. "Is faster r-cnn doing well for pedestrian detection?." In *European conference on computer vision*, pp. 443-457. Springer, Cham, 2016.

[34] Li, Jianan, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. "Scale-aware fast R-CNN for pedestrian detection." *IEEE transactions on Multimedia* 20, no. 4 (2017): 985-996.