CSSD: An End-to-End Deep Neural Network Approach to Pedestrian Detection

Feifan Wei, Jianbin Xie, Wei Yan, Peiqin Li

CSSD: An end-to-end deep neural network approach to pedestrian detection

Feifan Wei, Jianbin Xie, Wei Yan, Peiqin Li

School of Electronic Science, National University of Defense Technology,
Changsha, Hunan, 410073, P. R. China.
wff0316@foxmail.com

Abstract. SSD(Single Shot Multibox Detector) provides a powerful framework for detecting objects using a single deep neural network. The detection framework is one of the top object detection algorithms in both accuracy and speed which processes a large set of object locations sampled across an image. However, this framework does not behave well for the task of pedestrian detection since the images in popular pedestrian datasets have multiple objects occlusion problem and contain lots of small objects. In this paper, we incorporate deconvolution and downsampling unit into the SSD framework allowing detection network to recycle feature maps learned from images. The enhanced performance was obtained by changing the structure of classifier network, e.g., by replacing VGGNet with DenseNet. The contribution of this paper is a one-stage approach to compose a single deep neural network for pedestrian detection task in real-time. This approach addresses the typical difficulty of detecting different scale pedestrian at only one layer by providing a novel channel fusion. To solve small objects problem, base network has been replaced with more powerful one. This approach outperforms competing one-single methods on standard Caltech pedestrian dataset benchmark. It is also faster than all the other methods.

Keywords: deep learning, pedestrian detection, one-stage detector

1 INTRODUCTION

Pedestrian detection is one of main areas of researches in computer vision, due to its importance for a number of human-centric applications, such as video surveillance, autonomous driving, person identification and robotics[i]. Real-time accurate detection of pedestrians is a key for these systems.

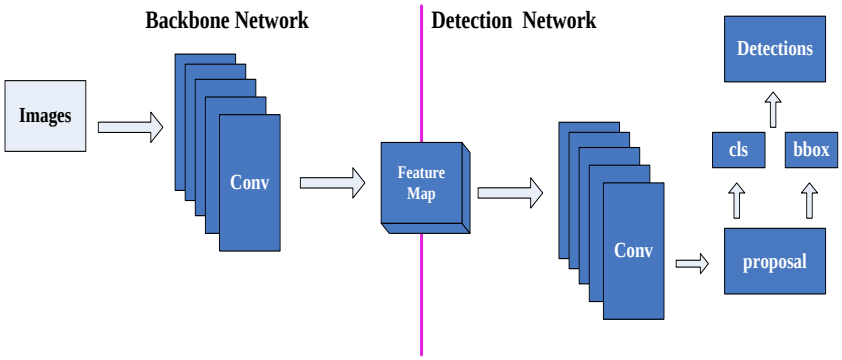
In recent years, convolutional neural networks (CNNs) have been applied to pedestrian detection algorithms in various ways, to improve the accuracy and speed of pedestrian detection[ii,iii,iv,v]. As illustrated in [vi], the detection algorithms based on deep learning can be divided into two-stage detectors and one-stage detectors, and one-stage detectors has the potential to achieve faster and better results. As shown in Fig. 1, one-stage detectors can be divided into two parts: backbone network and detection network.

Among various one-stage detection methods, SSD(Single Shot Multibox Detector)[vii] is one of the few algorithms that can guarantee robustness in real-time detection, because it adopts multiple convolution layers for detection. Although the conventional SSD performs well in object detection, it still has a few problems when applying it to pedestrian detection since the images in popular pedestrian datasets have multiple objects occlusion problem and contain numerous small objects.

Firstly, each layer in the feature pyramid of SSD is used independently as an input to the classifier network. Thus, the method can detect pedestrians of different scales in one picture, meanwhile the same pedestrian can be detected in multiple scales. However, SSD looks at only one layer for each scale, so it does not consider the relationships between the different scales. For example, in Fig.2, SSD finds two boxes for one person. Pedestrians in front of images tend to have higher confidence than later ones. After applying the NMS algorithm, the detection boxes of pedestrians behind are often suppressed. This is why the SSD algorithm does not perform well for occlusion problems.

Secondly, SSD is not robust to small-scale pedestrian detection. This problem is ubiquitous in current detection algorithms. In an image, small-scale pedestrians have smaller receptive fields and it is more difficult to extract their features. To solve this problem, there have been many attempts such as increasing the number of channels in one layer or replacing the basic network with more powerful one.

Fig. 1. . Overview of the proposed CSSD framework.



In this paper, these problems are tackled as follows: At first, the backbone network in the SSD is replaced with the improved DenseNet[viii]. As shown in Table 1, the network structure has been deepened , thereby improving feature extraction capabilities, with fewer parameters and faster convergence. Then, a circular feature pyramid is set up by deconvolution and downsampling units, after which the algorithm CSSD(Cycle Single Shot Detector) is named. The circular feature pyramid can make full use of the information between each layer to accurately predict the pedestrian detection boxes. Finally the algorithm performs better on occlusion than the others.

As shown in Fig. 1, the proposed detection algorithm can be divided into two parts: backbone network and detection network. DenseNet and CSSD are used as the backbone network and detection network, respectively.

The contributions made by this paper can be summarized as follows:

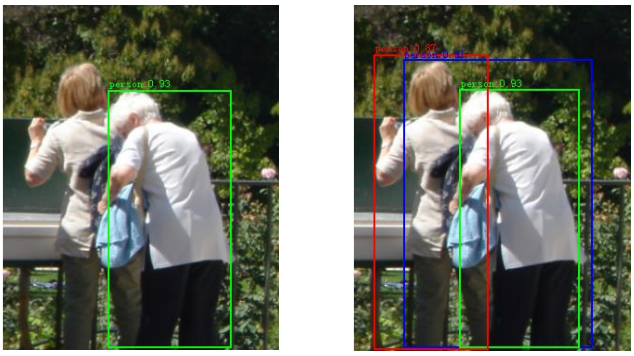
- The DenseNet network have been modified to improve the feature extraction capabilities of the backbone network and to enhance the ability of the algorithm to detect small-scale pedestrians.
- A circular feature pyramid has been designed to make full use of the information in each layer of the network to detect pedestrian boxes more accurately , thereby improving the algorithm's robustness to occlusion.
- The state-of-the-art performance has been achieved in the one-stage detectors on Caltech pedestrian dataset.

2 CSSD

In this section, CSSD(Cycle Single Shot Detector) is introduced. As shown in Fig. 1, the CSSD is divided into backbone network for extracting features and detection network for the generation and classification of candidate boxes.

We have improved SSDs in three ways:

- A brand-new backbone network is designed following the DenseNet framework.
- The deconvolution and downsampling modules are used to construct a circular feature pyramid in order to make full use of the information in each layer of the network.
- The number of channels in each layer is reduced with fewer parameters and higher efficiency.



(a) (b)

Fig. 2. Detection results of SSD, boxes with person score of 0.5 or higher is drawn: (a) SSD with two boxes for the right person. (b) after applying NMS, only the right person can be detected.

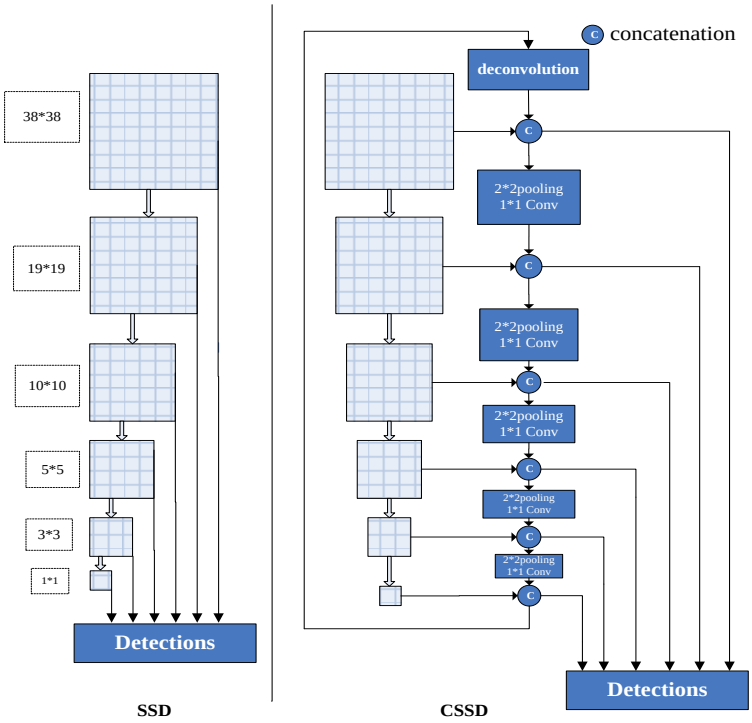


Figure 3. The comparison between SSD and CSSD detection network structure.

2.1 Backbone Network

Our base network is a variant of DenseNet. The network structure of the entire CSSD is shown in Table 1. The backbone network consists of one Stem block, four Dense blocks and four transition layers. Inspired by the Inception-v4[ix] network, we used three 3*3 convolutional layers and a 2*2 pooling layer are adopted to form the Stem block. The Stem block replaces the 7*7 convolutional layer in DenseNet, which increases the network depth and improves the network feature extraction capability while ensuring the same receptive field. By comparing with other classification networks, the performance of the network in the experimental part will be demonstrated.

2.2 Circular Feature Pyramid

In order to make full use of the information in each layer and integrate more high-level semantic information in detection, as shown in figure 3, a circular feature pyramid using down-sampling and deconvolution modules is set up. The CSSD model is improved from SSD with modified DenseNet. The number of channels in each layer is reduced during downsampling. The two reasons why a paradigm that predicts every layer is not used in the FPN[x] algorithm are as follows: On the one hand, pedestrian detection is a fundamental task in a system, which needs to provide

enough information for the downstream task. Therefore, speed is a key to the algorithm. Making predictions on each layer means the time for inference will increase several times. This is not acceptable for a fast pedestrian detection algorithm. On the other hand, a deconvolution module can reduce the number of parameters, improve efficiency, and make full use of information between layers. The more accurate detection boxes can be obtained, the less impact of occlusion will be on detection.

Table 1. CSSD backbone network

Layers		Output Size	Structure
Stem	Convolution	150×150	3×3 conv, stride 2
	Convolution	150×150	3×3 conv, stride2
	Convolution	150×150	3×3 conv, stride2
	Pooling	75×75	2×2avg pool, stride2
Dense Block (1)		75×75	$\begin{bmatrix} 1\times1\,\text{conv} \\ 3\times3\text{conv} \end{bmatrix}\times6$
Transition Layer (1)		75×75	1×1 conv
		38×38	2×2avg pool,stride 2
Dense Block (2)		38×38	$\begin{bmatrix} 1\times1\,\text{conv} \\ 3\times3\text{conv} \end{bmatrix}\times8$
Transition Layer (2)		38×38	1×1 conv
		19×19	2×2avg pool,stride 2
Dense Block (3)		19×19	$\begin{bmatrix} 1\times1\,\text{conv} \\ 3\times3\text{conv} \end{bmatrix}\times8$ $\begin{bmatrix} 1\times1\,\text{conv} \\ 3\times3\text{conv} \end{bmatrix}\times8$
Transition Layer (3)		19×19	1×1 conv
Dense Block (4)		19×19	$\begin{bmatrix} 1\times1\,\text{conv} \\ 3\times3\text{conv} \end{bmatrix}\times8$
Transition Layer (4)		19×19	1×1 conv

2.3 Connection Modules

As shown in Fig 3, there are two types of connection modules in the CSSD network. The first is that a module has been designed to combine the downsampling layer with a 1*1 convolution. The downsampling connection units has the following two roles:

- The upper feature size is processed by the downsampling unit to be consistent with the size of the underlying feature map;
- The channel fusion of the upper feature map keeps the number of output channels within a reasonable range.

The second type is , the deconvolution unit has been adopted to fuse the feature map of the last layer of the detection module with the feature map of the first layer. In this way the information flow of the network can be recycled. Each layer can obtain information from other layers. Similar to DSSD, the deconvolution unit consists of 3 convolution layers, 3 Batch Normalization layers, 1 deconvolution layer, 2 Relu activation functions, and a connection unit. The main difference is that we are connected to the first and last layer of feature pyramid.

2.4 Training Objective

The CSSD training objective is derived from the SSD[7] objective but just handles one object category. A set of prior bounding boxes at different scales and aspect ratios are generated at each position in the image. A default bounding box is labeled as positive if it has a Jaccard overlap greater than 0.5 with any ground truth bounding box, otherwise negative.

$$labels = \begin{cases} 0, otherwise \\ \\ 1, \text{ if } \dfrac{A_d \cap A_g}{A_d \cup A_g} > 0.5 \end{cases} \tag{1}$$

where A_d and A_g represent the default bounding box and the ground truth, respectively. The training objective is given as Equation(2):

$$L = \frac{1}{N}(\alpha L_{loc} + L_{conf}) \tag{2}$$

where N is the number of matched default boxes, and α is a constant weight term to keep a balance between the two losses. L_{conf} is the softmax loss over person category confidence.

$$L_{conf} = -\sum_{i \in Pos}^{N} x_{ij} \log(c_i) - \sum_{i \in Neg} \log(c_i)$$

$$\text{where } c_i = \frac{\exp(c_i)}{\sum_p \exp(c_i)} \tag{3}$$

where Pos and Neg represent the positive and negative default boxes, respectively. $x_{ij} = \{1,0\}$ is an indicator for matching the i-th default box with the j-th ground truth box of category person. L_{loc} is the Smooth L1 loss[xi], not modified, for more details about the loss please refer to[7].

3 EXPERIMENTS

3.1 Dataset and Evaluation

The Caltech pedestrian dataset is currently the most commonly adopted and challenging dataset in pedestrian detection. The pedestrian detection algorithm proposed in this paper was evaluated on the Caltech dataset. A total of approximately 250,000 frames 350,000 rectangular frames and 2300 pedestrians were marked. The original frame size is 480×640. The FPPI standard is proposed by the Caltech dataset to evaluate the algorithm.

3.2 Training , Testing Settings and Results

The pedestrian detector proposed in this paper is implemented in the caffe framework. Most training strategies follow the SSD algorithm, including data augment and loss function settings. Due to the difference between the scale ratio of the pedestrian and the common object in the object recognition, the scale setting of the prior box has been adjusted in the algorithm. In addition, because the algorithm is trained from scratch, the setting of hyperparameters such as learning rate will also be different. This algorithm is trained on **NVIDIA 1080Ti GPU**.

The Caltech datasets includes a total of 11 video segments (s0-s10), of which the first 6 video segments are training sets and the last 5 video segments are used as testing sets. The original size of these images is 480*640. In the setting of the candidate box and aspect ratio, the method of F-DNN has been basically followed. Experiments show that more candidates can be obtained through the method. Therefore the matches to the ground truth are not lost. By using the SGD training method, a batchsize of 2, the learning rate of 10^{-5}, and all weights are randomly initialized and trained from scratch.

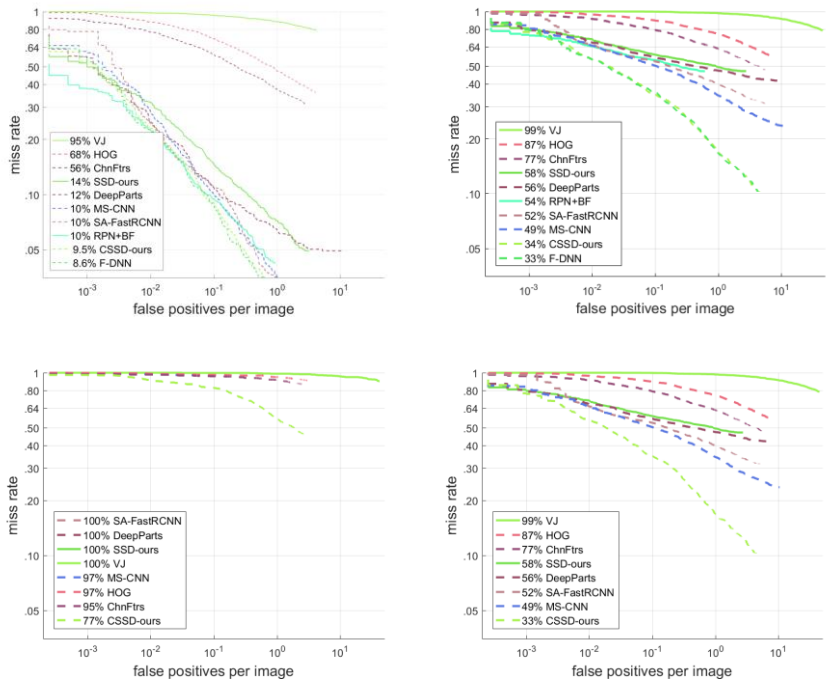


Fig. 3. Comparison of CSSD with the state-of-the-art methods on the Caltech dataset using the reasonable, occ.partial setting, far and medium setting.

CSSD has achieved an impressive 9.5% missing rate on the Caltech dataset using reasonable setting. Compared with SSD-ours, this algorithm achieved a 32% improvement. SSD-ours is a learning algorithm based on SSD framework for pedestrians. As shown in Fig 3, the ROC plot of missing rate against FPPI is shown for the current top performing methods reported on Caltech. To the best of our knowledge, this detector is the first true one-stage pedestrian detector, and has achieved similar performance to the state-of-the-art algorithm. Not surprisingly, compared with SSD-ours, this algorithm achieved a 56% improvement using the Occ.partial setting and 77% improvement using far setting. In other words, our algorithm performs extremely well in dealing with pedestrian occlusion and small-scale pedestrian.

4 RESULTS ANALYSIS

4.1 Effectiveness Analysis

Through ablation experiments, the role of backbone network and detection network has been explored in the CSSD network. Figure 5 visualizes the results of SSD and our method. At the beginning, we used the original SSD algorithm to train and test on the Caltech dataset, achieving a 14.3% missing rate in the reasonable setting. After replacing the backbone network with DenseNet, the performance of the algorithm has

improved from 14.3% to 11.6%. After the backbone network employed VGG, and the detection structure has been replaced with the proposed CSSD, the performance of the algorithm has improved from 14.3% to12.4%. Having adopted DenseNet and the detection network CSSD, the miss rate finally reached 9.56%, as shown in Table 2.

Table 2. Effectiveness of the backbone network and detection network

Method	Reasonable
VGG+SSD	14.3%
DenseNet+SSD	11.6%
VGG+CSSD	12.4%
DenseNet+CSSD	9.50%

4.2 Runtime Analysis

As shown in table 3, compared with the recent state-of-the-art algorithms, our algorithm framework uses only a single convolutional framework. Thus the processing speed of an image is only 0.08s, which completely meets the needs of real-time processing. In the currently known algorithms, no other pedestrian detection algorithm can achieve the processing speed achieved by the CSSD algorithm with the same accuracy. This algorithm is designed for pedestrian detection in real-world scenarios. Since it is possible to train from scratch, the CSSD algorithm has a very strong migration capability and can be used in a very wide range of applications. This is one of the advantages of the CSSD algorithm.

Table 3. A comparison of speed among the state-of-the-art models

Method	Caltech(%)	Runtime（GPU）
DeepParts	11.89	1s
ComACTr-Deep	11.75	1s
MS-CNN	9.95	0.4s
SA-FastRCNN	9.68	0.59s
RPN+BF	9.58	0.60s
F-DNN	8.65	0.30s
SSD	13..06	0.06s
CSSD(ours)	**9.50**	**0.08s**

5 CONCLUSIONS

An efficient and robust one-stage pedestrian detector has been proposed based on a single DNN trained from scratch. A brand new network CSSD for pedestrian detection is designed. To the best of our knowledge, this algorithm is state-of-the-art one-stage pedestrian detector on Caltech datasets. Furthermore, CSSD has great potential on special domain scenario like military district early warning, night guard, etc.

For future work, the pedestrian detection system based on semantic segmentation has achieved the best detection results. Due to the versatility of the CSSD network, semantic segmentation modules can be easily integrated into the network and will achieve better results. This part of the work will be the focus of our future research.

References

1. A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In Conference on Computer Vision and Pattern Recognition(CVPR), 2012. 1, 2, 5, 6
2. J. H. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians. CoRR, abs/1501.05790, 2015.
3. X. Du, M. El-Khamy, J. Lee, and L. S. Davis. Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection. arXiv preprint arXiv:1610.03466, 2016. 2, 3, 6
4. Y. Tian, P. Luo, X.Wang, and X. Tang. Deep learning strong parts for pedestrian detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 1904-1912, 2015. 2, 6
5. L. Zhang, L. Lin, X. Liang, and K. He. Is faster R-CNN doing well for pedestrian detection? To appear in ECCV 2016, 2016.
6. TsungYi Lin, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection[J]. 2017:2999-3007.
7. Wei Liu, 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, pages 21–37. Springer, 2016.
8. G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In CVPR, 2017.1, 2, 3, 4
9. Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[J]. 2016.
10. Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." CVPR, 2017.
11. R. Girshick. Fast R-CNN. In International Conference on Computer Vision (ICCV), 2015.