Fast Pedestrian Detection With Attention-Enhanced Multi-Scale RPN and Soft-Cascaded Decision Trees

Han Wang, Yali Li, *Member, IEEE*, and Shengjin Wang, *Senior Member, IEEE*

***Abstract*— Pedestrian detection has attracted more attention in the fields of computer vision and artificial intelligence. A variety of real-world applications involving pedestrian detection have been promoted, such as Advanced Driving Assistant System (ADAS). Although both two-stage and single-stage deeply learned object detectors have shown outstanding performance for general object detection, they are still facing the problem of poor accuracy in single-class detection senario because they are designed to distinguish objects from different categories rather than pay attention to various appearances of pedestrians. Previous leading pedestrian detectors F-DNN and F-DNN v2 fuse several neural networks like SSD, VGG16 and GoogLeNet to generate ROIs and supress false alarms with cascaded structure, resulting in low miss rate but high complexity. In this paper we propose a novel framework called Attention-Enhanced Multi-Scale Region Proposal Network (AEMS-RPN) for ROI generation, which also acts as first-stage classification. Inspired by the success of traditional pedestrian detectors, we use soft-cascaded decision trees instead of cascaded deep neural networks to achieve high accuracy and fast detection speed simultaneously. The decision tree classifier is used and enables us to combine features from different layers with various resolutions for classification and incorporate effective bootstrapping for mining hard negatives. We test our method on several pedestrian detection datasets and the experimental results certify the effectiveness of the proposed AEMS-RPN. Compared with the state-of-the-art, we obtain the competitive accuracy with near real-time efficiency.**

***Index Terms*— Pedestrian detection, attention mechanism, multi-scale, soft-cascade, DNN.**

I. Introduction

PEDESTRIAN detection has attracted more and more attention because of its significant role in real-world artificial intelligence applications, such as intelligent surveillance, ADAS and automatic driving. Fig. 1 shows several pedestrian detection results along with our enhanced features on street scenes.

Manuscript received February 25, 2019; revised October 6, 2019; accepted October 15, 2019. Date of publication October 25, 2019; date of current version November 30, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61701277 and Grant 61771288 and in part by the State Key Development Program in 13th Five-Year under Grant 2017YFC0821601. The Associate Editor for this article was S. S. Nedevschi. *(Corresponding author: Yali Li.)*

The authors are with the Institute for Artificial Intelligence, Tsinghua University (THUAI), Beijing 100084, China, with the Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China, and also with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: han-wang@outlook.com; liyali13@mail.tsinghua.edu.cn; wgsgj@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TITS.2019.2948398

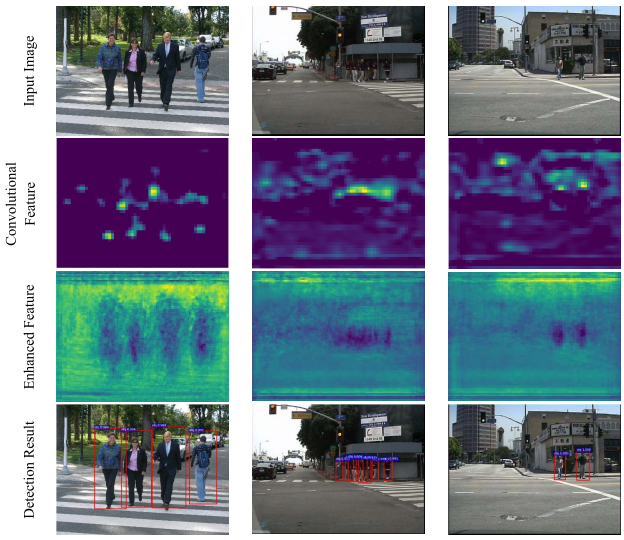


Fig. 1. Examples of Enhanced Features and Pedestrian Detection Results on Street Scenes.

Typically, a pedestrian detection process consists of three steps:

Region Proposal Pre-process the input image and select regions where pedestrians may exists.

Feature Extraction Compute fixed-length features for each proposed region.

Binary Classification Feed features into a binary classifier and judge whether a pedestrian exists.

Hand-crafted features appear to have an significant role in leading pedestrian detectors because of its higher resolution in contrary to convolutional features. Pedestrian instances in automatic driving, intelligent surveillance and other typical scenarios are generally of small sizes (e.g., 28×70 for Caltech-USA [1]), while the computed convolutional feature map in Faster R-CNN has a stride of 16 pixels [2]. The low resolution of features limits their discrimination ability, and furthermore degrades the subsequent classifier.

Reference [3] argues that Multi-Layer Perceptron (MLP) fails to pay attention to hard negative examples and its detection accuracy is even worse than initial region proposals. Cascaded Boosted Forest (BF) [4] attached to Region Proposal Network(RPN) with bootstrapping strategy achieves a better performance. However, their detection speed is 500ms per image, which is too slow for real-world applications. Another top-performance approach Fused-DNN combines SSD [5], VGG16 [6], GoogLeNet [7] producing models larger than

700MB for pedestrian detection. The complexity of models makes their algorithm unrealistic in portable devices and real-world applications such as intelligent surveillance and automatic driving.

We focus on to achieve both high accuracy and fast speed in pedestrian detection within one model. In this paper we propose a framework of Attention-Enhanced Multi-Scale RPN (AEMS-RPN) focusing on improved region proposals to handle large scale variance and low confidence targets caused by intensity, blurring or occlusion problems. Specifically, we propose multiple region proposal subnetworks to handle large scale variance of pedestrians. Large separable convolution [8] is adopted to build thinner and more robust feature maps. Considering the information loss in deeper convolution layers, we introduce an attention mechanism to enhance high-level pedestrian features. We find that an active path in MLP classifier is equivalent to a decision tree and combining multiple decision trees for inference requires much less parameters. Experiments on Caltech-USA pedestrian detection benchmark show that our approach achieves the state-of-the-art performance with fast speed.

The main contribution of this work can be summarized as: (**I**) We propose a framework Attention-Enhanced Multi-Scale RPN to handle large-scale variance of pedestrians and suppress false positives including double detections, body parts and background clutters. (**II**) Large separable convolution is introduced to provide rich context and a novel attention mechanism is incorporated to overcome the drawbacks of large receptive field, requiring no additional annotations (e.g. segmentation, optical flow). (**III**) Using soft-cascaded decision trees instead of a multi-layer perceptron (MLP), our method achieves the state-of-the-art accuracy with a compact model and near real-time inference speed.

II. Related Works

A. Region Proposal

The simplest and most effective method of region proposal is sliding window, which resizes the input image to different scales and applies a fixed-size sliding window to each scale generating candidate regions. With the development of neural network-based object detection algorithms, Faster R-CNN with RPN [2] becomes the leading two-stage, proposal-driven detection mechanism. RPN is actually a sliding window method built on convolutional features with a cascaded binary classifier.

B. Feature Extraction

Hand-crafted features for pedestrian detection have been studied for more than a decade. Viola and Jones proposed the first practical pedestrian feature representation in 2005 [9] and their contribution has become a fundamental algorithm in OpenCV, a famous computer vision toolkit, now. Driven by the hypothesis that the diversity of features drawn from the input image can improve detection quality, researchers have explored numerous feature categories since then: edge information [10]–[13], color information [11], texture information [14], local shape information [15], amongst others.

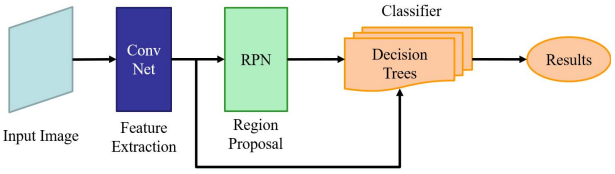


Fig. 2. Our Proposed Pipeline: The convolutional features are shared by the RPN and the downstream classifier.

With the aid of forward-backward propagation algorithm, we are now able to learn deep convolutional features for pedestrian representation via deep neural networks [6], [16]–[18]. Due to the low resolution of deep convolutional features, some leading pedestrian detectors (e.g., [17]–[19]) combine traditional hand-crafted features and prevalent convolutional features hoping to achieve a balance between resolution and discrimination.

C. Classifier

In traditional pedestrian detection approaches, most hand-crafted features are classified by Support Vector Machines(SVMs) or Decision Trees. Rodrigo Benenson and Mohamed Omran et al. argue that it is not sufficient to draw conclusion whether SVMs or Decision Trees are better classifiers for pedestrian detection tasks [20]. With the quick development of deep learning, a typical neural network for pedestrian detection includes convolution layers followed by fully-connected layers. Fully-connected layers forms a MLP that can solve non-linear classification problems. Fully-connected layers will slow down the detection speed due to the huge quantities of parameters and, even worse, degrade region proposal results as suggested in [3].

In summary, the major obstacle for deploy existing object detection approaches in pedestrian detection application is their poor representation of feature maps for pedestrian-specific region proposal. For inference accuracy and time efficiency, soft-cascaded decision trees are needed to be adopted as classifier. In this paper, we design a better region proposal structure called Attention-Enhanced Multi-Scale RPN (AEMS-RPN), attached with soft-cascaded decision trees for classification.

III. Our Approach

The proposed approach consists of three components (illustrated in Fig. 2): the ConvNet for convolutional features, the RPN for candidate proposals, and the soft-cascaded decision trees that classify these proposals based on convolutional features.

A. Region Proposal

RPN in Faster R-CNN [2] originally acts as a foreground/background single-class detector in the first-stage of a two-stage multi-class detector. For single-class task like pedestrian detection, two-stage detectors are less efficient since RPN itself is actually a high-efficiency one-stage detector.

We modify the original RPN to be better fit for pedestrian detection tasks. Following [3], we employ anchors of 0.41 aspect ratio (an average aspect ratio of pedestrian targets)

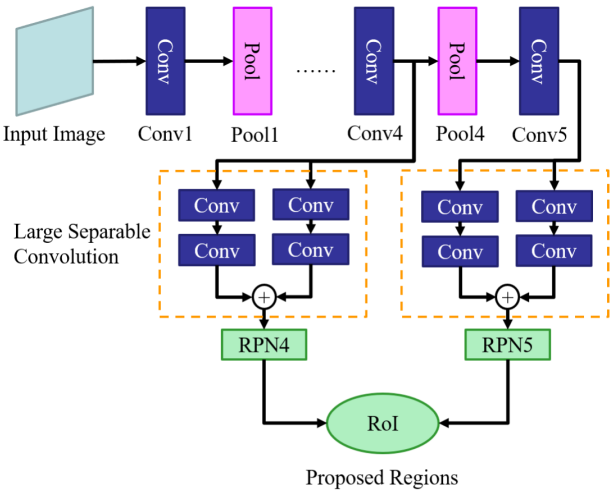


Fig. 3. Multi-Scale Region Proposal.

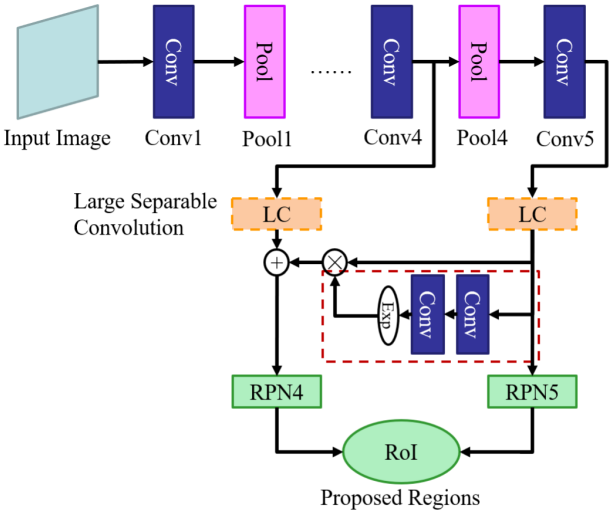


Fig. 4. Attention Mechanism.

rather than three aspect ratios $\{0.5, 1, 2\}$ suggested in [2]. In addition, we use 9 anchors with a scaling stride of $1.3\times$, starting from $40px$ height.

Duplicate detections and body parts are the most common false positives in pedestrian detection and are more difficult to suppress comparing with other non-human targets. The feature activations relating to body parts are consistent with our learning objective and a larger context will help the network produce full-body pedestrian targets rather than separate parts. In a convolution layer, kernel size k determines the receptive field of a neuron and increasing k is the easiest way to provide a larger context. $k \times k$ convolution with large k is time-consuming and we adopt large separable convolution [8] to convert it into one $k\times1$ convolution follow by $1\times k$ convolution and one $1 \times k$ convolution followed by $k \times 1$ convolution. The numbers of output channels are denoted as c_{mid}, c_{out}.

Another major difficulty of pedestrian detection is large scale variance. The original work believes that pooling features of each ROI into the same size can eliminate the necessity of feature pyramid. Unfortunately, the process can also cause information loss for large-scale targets and produce insufficient features for small-scale ones. In this paper, we suggest two region proposal units working on different convolution layers, dealing with small and large pedestrian targets, respectively (Fig. 3). Two units regress boxes with different strides and meanwhile their classification layers provide confidence scores of the predicted boxes, which are mapped to a larger range and function as the initial scores of the soft-cascaded structure.

B. Attention and Feature Fusion

It is a common knowledge that deeper convolution layers lead to lower resolution while shallower ones lead to weaker representation. In multi-scale RPN, proposals are generated from convolution layers with different strides and also different representation ability. Detecting smaller targets require feature maps with higher resolution and equivalent feature representation, which can be achieved by directly fuse feature maps of two different layers with up-sample and sum operations as shown in Fig. 4 (except the dashed box).

Visual attention mechanism is always effective in object detection methods, in that it could suppress the confusion by negative targets. Since our receptive field has been enlarged through large separable convolution, more background information is incorporated into the feature maps. The direct fusion without attention mechanism means that each pixel in high-level feature map has equal contribution to low-level features and the background clutters brought to low-level feature map have severer influence on smaller targets. Improvement can be made by introducing attention mechanism and emphasizing on features related to pedestrian targets. Without segmentation annotations, we generate human masks in our network structure with two convolution layers and one exponential operation layer. Feature map $f_h(x)$ from higher level is firstly up-sampled to match the shape of lower layer's output, then two 3×3 convolution layers are applied to generate an exponential mask $\sigma(f_h^{up}(x))$. The mask is later applied to $f_h(x)$ and the final feature map is computed according to Eq. 1, where $f_l(x)$ denotes feature map from the lower layer (the dashed box in Fig. 4).

$$f_m(x) = f_l(x) + f_h^{up}(x) \cdot \exp\left(\sigma\left(f_h^{up}(x)\right)\right) \tag{1}$$

The exponential mask has a similar function during back-propagation (Eq. 2) which provides a positive feedback on pixels related to human targets in the attention mask. This mask maps zeros to ones and enlarges non-zero values and will highlight feature activations related to true pedestrian targets when multiplied to feature maps according to Equation 1.

$$\frac{\partial L}{\partial \sigma} = \frac{\partial L}{\partial f_m} \cdot \frac{\partial f_m}{\partial \sigma\left(f_h^{up}\right)} = \frac{\partial L}{\partial f_m} \cdot \left(f_h^{up} \exp\left(\sigma\left(f_h^{up}\right)\right)\right) \tag{2}$$

C. Feature Extraction

Extracting accurate features for small-scale targets is important in pedestrian detection. The original Faster R-CNN quantize the location of ROIs into integers on feature maps, causing a misalignment up to 30 pixels on input images. This has a significant negative influence on pedestrian detection. We use ROI-Align [21] with bilinear interpolation to draw

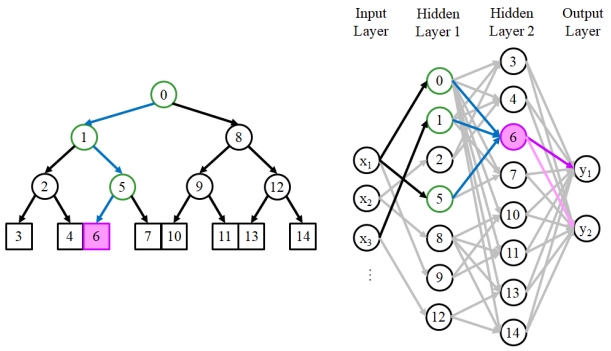


Fig. 5. Decision Tree VS. MLP Classifier.

features of fixed length from each region. These features will be used to train decision trees as introduced in the next section.

Since decision tree classifier essentially minimizing the loss by greedy search, its input dimension can be arbitrary. We make full use of the features extracted from ROIs on different convolution layers. We pool these features into a fixed resolution and simply concatenate the features without normalization. In contrast, a delicate feature normalization is needed when concatenating features for MLP classifiers [22].

The feature of small-scale pedestrians may become trivial if their size is close to or even smaller than original convolution strides. In this case, we use átrous trick [23] to dilate convolution kernels, generating feature maps of higher resolution given the fine-tuned layers from Attention-Enhanced Multi-Scale RPN. For example, considering a group of convolution layers, we shrink the window size of the previous pooling operation by 2 firstly and then dilate all convolution kernels in the group by 2. In this way, the stride of output feature map is reduced by half.

D. Soft-Cascaded Decision Trees

While the leading detection algorithms use fully-connected layers for classification [2], [8], we argue that decision trees have the same classification ability as MLP classifier composed with fully-connected layers. As Fig. 5 shows, an activate path in MLP classifier involving three L1-hidden neurons and one L2-hidden neuron is equivalent to a path in decision tree. The decision tree consists of three decision nodes and one prediction node where decision nodes act as ReLU activation in MLP and prediction nodes act as Soft-Max classification. Combining multiple decision trees with effective bootstrapping method can achieve stronger classification ability than MLP. Another advantage of decision trees is that only $\log_2 N$ parameters are used during inference in contrast with N parameters in MLP. In our soft-cascaded structure, if some decision trees are firmly convinced that an ROI is negative, the rests are not likely to classify the same ROI as positive. If an ROI scoring lower than a threshold, for example -1, it can be pruned in advance to further improve the inference speed and interests some recent works [24], [25].

Attention-Enhanced Multi-Scale RPN generates the region proposals, confidence scores and features required for training soft-cascaded decision trees by RealBoost algorithm [26]. All positive examples from annotations and a similar quantity of randomly sampled negatives constitute the training set. The training process is bootstrapped by several times and some extra top-rated hard negative examples (10% of the positive number) are attached to the training set after each stage. Decision trees in the final stage are used for inference in a soft-cascaded structure.

The confidence scores given by Attention-Enhanced Multi-Scale RPN is in the range of $[0, 1]$, while those of decision trees are much larger. In order to accumulate scores from different sources, we map the score s given by RPN to f_0 using:

$$f_0 = \frac{1}{2}\log\left(\frac{s}{1-s}\right) \tag{3}$$

f_0 and scores $f_i, i = 1, 2, \cdots, 2048$ given by each decision trees add up together to give the final assessment for predicting whether a targeted region is a pedestrian.

E. Implementation Details

We use VGG16 [6] as the backbone of our Attention-Enhanced Multi-Scale RPN. Large separable convolution units with kernel size $k = 7$, medium output channel $c_{mid} = 256$, final output channel $c_{out} = 128$ and kernel size $k = 5$, medium output channel $c_{mid} = 256$, final output channel $c_{out} = 128$ are attached to layer Conv4_3 and layer Conv5_3 separately, and followed by two region proposal units. Each unit contains a 3×3 convolution layer and two parallel 1×1 convolution layers for classification and bounding box regression. Other hyper-parameters of Attention-Enhanced Multi-Scale RPN are the same as in [2]. With proposed regions, convolutional features are extracted from different combinations of convolution layers and pooled into a fixed-size of 7×7.

We employ image-centric training and testing as in [2]. An input image is resized and its shorter edge has 720. For the training of Attention-Enhanced Multi-Scale RPN, anchor boxes with an Intersection-over-Union (IoU) ratio larger than 0.5 with ground-truth are considered positive and the rest are negative. Each mini-batch for computing the loss consists of 1 image and 120 randomly sampled anchors. As addressed in [3], cross-boundary negative anchors are preserved to improve accuracy on Caltech-USA dataset.

With the fine-tuned Attention-Enhanced Multi-Scale RPN, the proposal regions are filtered by Non-Maximum Suppression (NMS) with a threshold 0.7 and then ranked by their scores. The following classifier is composed of decision trees with depth 5. While training the decision trees, we accumulate 1000 top-rated proposals of each image into the training set. While performing inference, we only keep 100 top-rated proposals to enhance detection speed.

IV. Experiments and Analysis

A. Dataset

We evaluate our approach on Caltech-USA [1] and by default an IoU threshold 0.5 is used. Following [19], we augment the training data by 10 folds (resulting in 42782 images in total) and evaluate our proposed approach on the standard

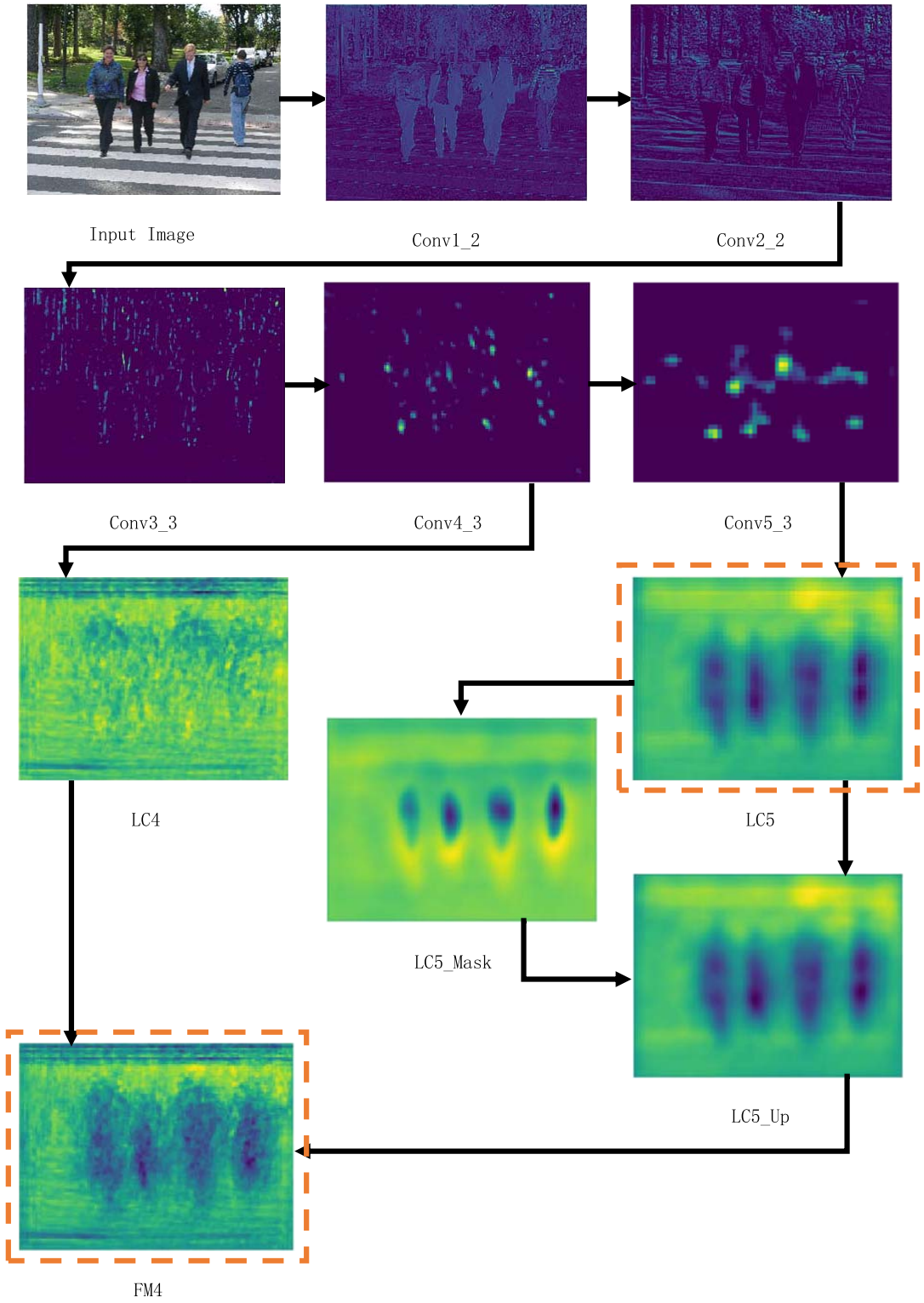


Fig. 6. Visualization of Attention-Enhanced Multi-Scale RPN.

test set under the reasonable settings (pedestrian targets lower than 50 *px* or less than 65% visible are dropped). We report log-average Miss Rate on False Positive Per Image (FPPI) in the range of $[10^{-2}, 100]$ (denoted as MR in short) as evaluation metric. Further experiments are conducted on CityPersons [27], a new pedestrian detection dataset based on the semantic segmentation dataset CityScapes [28].

B. Attention-Enhanced Multi-Scale RPN

Firstly, we conduct an experiment to compare the proposal quality of different region proposal methods. Feed an input image into Attention-Enhanced Multi-Scale RPN, and the outputs of each major layer are visualized in Fig. 6. The first group of convolution layers act as area segmentation,

TABLE I
COMPARISON OF DIFFERENT REGION PROPOSAL METHODS ON CALTECH-USA

Method	Multi-Scale	Large Separable Convolution	Feature Fusion	Attention	MR(%)
RPN[3]					14.90
RPN+	√				14.07
RPN+	√	√			13.23
RPN+	√	√	√		12.68
RPN+		√	√	√	12.71
AEMS-RPN	√	√	√	√	12.17

the second group as edge extraction and the third as corner and vertical structure detection. In the fourth and fifth group of convolution layers which serve as feature maps for classification in many other pedestrian detectors [3], [29], [30], locations where pedestrians appear are activated. However, we argue that these feature maps contain too many channels and zero values for following classifier. To solve this, a large separable convolution [8] can not only reduce computation but also make features more powerful due to larger valid receptive field caused by large kernel. Our experiment shows that LC4 and LC5 outputs by large separable convolution units have fewer channels but more non-zero values and the locations of each pedestrian targets are clearly visible. Attention mask LC5_Mask is generated from LC5 and later applied to LC5 to form LC5_Up with more details. LC4 and LC5_Up are added together generating FM4. Two region proposal subnets are attached to FM4 and LC5 (shown in orange dashed boxes) and the discriminative feature maps guarantee satisfactory results.

In order to assess the contribution of each modules, we add one module a time to original VGG16-based RPN. The proposal miss rates (MR) are listed in Table I for comparison. The original VGG16-based RPN achieve an MR of 14.9% on Caltech-USA and each component we proposed shows steady reduction in MR. Multi-scale RPN is capable of dealing with large scale variance; large separable convolution provides more context information to prevent double detections and body part detections; feature fusion ensures that feature maps associated to different scales have equivalent representation ability; attention mechanism overcomes the drawbacks of large receptive fields and brings a significant improvement. Our proposed attention-enhanced multi-scale RPN achieves an MR of 12.17%.

C. Features

Decision tree classifier is flexible and has no need of feature normalization, so the features for classification can be drawn from an arbitrary combination of convolution layers with no extra cost. Table II shows the results with different features fed to soft-cascaded decision trees in our method.

The decision trees are trained with bootstrapping strategy and each stage has a forest of {64, 128, 256, 512, 1024, 1536, 2048} trees. The bootstrapping strategy brings remarkable improvement in every situations and proven effective in pedestrian detection. We evaluate our proposed approach under reasonable settings in Caltech-USA benchmark and achieve 7.94% MR with original annotations and 6.02% MR with refined annotations by [25] (Fig. 7).

TABLE II
PEDESTRIAN DETECTION RESULTS WITH DIFFERENT FEATURES ON CALTECH-USA

Method	Features	MR(%)
AEMS-RPN + Decision Trees	Conv3_3	8.73
	Conv4_3	10.95
	Conv5_3	9.07
	Conv3_3, Conv4_3	10.70
	Conv3_3, Conv5_3	11.22
	Conv3_3, Conv4_3, Conv5_3	10.46
	Conv3_3, Conv4_3 with á trous	7.94
	Conv3_3, Conv5_3 with á trous	10.48

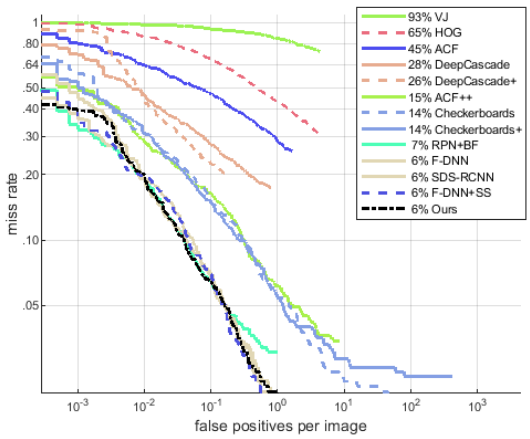


Fig. 7. Log-Average Miss-Rate on Caltech-USA with Refined Annotations.

D. Comparison

We achieve the state-of-the-art performance of 6.02% MR on Caltech-USA with refined annotations and the size of our model is only 103.4MB. RPN-BF [3] only achieves an MR of 6.81% due to the lack of multi-scale region proposal and attention mechanism. F-DNN-SS [31] has a similar MR comparing with our work but it requires a model > 700MB. Our approach runs at 10FPS on 720p input video stream with one NVIDIA TITAN X (Maxwell) GPU, which is near real-time. In contrast, most top-performance pedestrian detection methods run slower than 1FPS and limits their deployment in real-world applications.

We also evaluate our proposed framework on new CityPersons dataset in terms of reasonable evaluation setup. All models are trained on the train set and tested on the validation set and the results are shown in Table IV. Scaling-up 1.3× shows an immediate gain in the enhancement of small-scale pedestrian detection in the baseline method [28]. Large receptive fields cooperating with attention mechanism in our

TABLE III
COMPARISON WITH PREVIOUSLY LEADING DETECTORS ON CALTECH-USA

Detectors	Model Size (MB)	MR(%)
RPN+BF[3]	87.6	6.81
F-DNN[31]	241.8	6.31
F-DNN-SS[31]	769.8	6.04
Ours	103.4	6.02

TABLE IV
PEDESTRIAN DETECTION RESULTS ON CITYPERSONS

Method	Scale	MR(%)
Zhang, et al.[28]	$\times 1$	15.4
Zhang, et al.[28]	$\times 1.3$	12.8
Ours	$\times 1$	13.7
Ours	$\times 1.3$	12.2

work maintain a gap of 0.4% in MR with scaling-up, and a larger gap of 1.7% without scaling-up.

V. CONCLUSION

In this paper, we present a high-efficiency pedestrian detection method combining Attention-Enhanced Multi-Scale RPN (AEMS-RPN) and Soft-Cascaded Decision Trees. The former fuses features from different layers with the help of attention mechanism and makes region proposals of different scales. On top of region proposals and features pooled by ROI-Align, the soft-cascaded decision tree classifier is introduced to classify features of arbitrary resolutions and mine hard negatives through bootstrapping. These trees are soft-cascaded for pruning. This framework overcomes the defects of directly using Faster R-CNN systems for pedestrian detection and achieves the state-of-the-art accuracy with a compact model and near real-time inference speed.

REFERENCES

[1] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[3] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 443–457.

[4] R. Appel, T. Fuchs, P. Dollár, and P. Perona, "Quickly boosting decision trees: Pruning underachieving features early," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 594–602.

[5] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[7] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[8] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-head R-CNN: In defense of two-stage object detector," 2017, *arXiv:1711.07264*. [Online]. Available: https://arxiv.org/abs/1711.07264

[9] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, 2005.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[11] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.* London, U.K.: BMVC Press, 2009, pp. 91.1–91.11.

[12] J. J. Lim, P. Dollar, and C. L. Zitnick, III, "Learned mid-level representation for contour and object detection," U.S. Patent 13 794 857. Mar. 12, 2013

[13] P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable deep network for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 899–906.

[14] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 32–39.

[15] A. D. Costea and S. Nedevschi, "Word channel based multiscale pedestrian detection without image resizing and using only one classifier," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2393–2400.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[17] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1904–1912.

[18] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3361–3369.

[19] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4073–4082.

[20] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" 2014, *arXiv:1411.4304*. [Online]. Available: https://arxiv.org/abs/1411.4304

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[22] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," 2015, *arXiv:1506.04579*. [Online]. Available: https://arxiv.org/abs/1506.04579

[23] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[24] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 82–90.

[25] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1259–1267.

[26] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors)," *Ann. Statist.*, vol. 28, no. 2, pp. 337–407, 2000.

[27] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3213–3221.

[28] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[29] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 354–370.

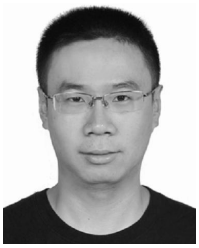
[30] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.

[31] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2017, pp. 953–961.

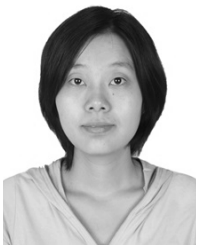
[32] Y. Li, S. Wang, Q. Tian, and X. Ding, "Feature representation for statistical-learning-based object detection: A review," *Pattern Recognit.*, vol. 48, pp. 3542–3559, Sep. 2015.

[33] Y. Li, S. Wang, Q. Tian, and X. Ding, "A survey of recent advances in visual feature detection," *Neurocomputing*, vol. 149, pp. 736–751, Feb. 2015.

[34] Y. Li, S. Wang, Q. Tian, and X. Ding, "A boosting approach to exploit instance correlations for multi-instance classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2740–2747, Dec. 2016.



Han Wang received the B.E. degree from Tsinghua University, Beijing, China, in 2017, where he is currently pursuing the Ph.D. degree in information and communication engineering with the Department of Electronic Engineering. His research interests include image processing, pattern recognition, computer vision, and machine learning. He was a recipient of the Excellent Graduates Award for his B.E. degree.



Yali Li received the B.E. degree from Nanjing University, China, in 2007, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2013. She is currently a Research Assistant with the Department of Electronic Engineering, Tsinghua University. Her research interests include image processing, pattern recognition, computer vision, and video analysis. She was a recipient of the Excellent Graduates Award for her B.E. degree.

Shengjin Wang (SM'19) received the B.E. degree from Tsinghua University, Beijing, China, in 1985, and the Ph.D. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 1997. From May 1997 to August 2003, he was a member of the Senior Research Staff with Internet System Research Laboratories, NEC Corporation, Nara, Japan. Since September 2003, he has been a Professor with the Department of Electronic Engineering, Tsinghua University, where he is currently a Director of the Research Center for Media Big-data Cognitive Computing. He has published more than 100 articles and possessed more than 20 patents. His current research interests include computer vision, pattern recognition, object detection, advanced driving assistant system, and deep learning.