Scale-adaptive Deconvolutional Regression Network for Pedestrian Detection

Yousong Zhu[1,2], Jinqiao Wang[1,2], Chaoyang Zhao[1,2],
Haiyun Guo[1,2], Hanqing Lu[1,2]

[1]National Laboratory of Pattern Recognition, Institute of Automation
[2]University of Chinese Academy of Sciences
{yousong.zhu, jqwang, chaoyang.zhao, haiyun.guo, luhq}@nlpr.ia.ac.cn

Abstract. Although the Region-based Convolutional Neural Network (R-CNN) families have shown promising results for object detection, they still face great challenges for task-specific detection, *e.g.*, pedestrian detection, the current difficulties of which mainly lie in the large scale variations of pedestrians and insufficient discriminative power of pedestrian features. To overcome these difficulties, we propose a novel Scale-Adaptive Deconvolutional Regression (SADR) network in this paper. Specifically, the proposed network can effectively detect pedestrians of various scales by flexibly choosing which feature layer to regress object locations according to the height of pedestrians, thus improving the detection accuracy significantly. Furthermore, considering CNN can abstract different semantic-level features from different layers, we fuse features from multiple layers to provide both local characteristics and global semantic information of the object for final pedestrian classification, which improves the discriminative power of pedestrian features and boosts the detection performance further. Extensive experiments have verified the effectiveness of our proposed approach, which achieves the state-of-the-art log-average miss rate (MR) of 6.94% on the revised Caltech [1] and a competitive result on KITTI.

1 Introduction

Pedestrian detection aims to locate all pedestrian instances with various poses, scales and occlusions in an image. Over the last decade, it has become a hot topic for its wide applications, such as smart vehicles, video surveillance and robotics.

Recently, a lot of efforts have been devoted to pedestrian detection based on boosted decision forests [2–6], deformable part model [7–9], and deep learning models [10–12]. No matter what kinds of methods, the scale of pedestrian which largely affects the feature representation or even dominates the detection performance still has not been well solved. In traditional methods, sliding windows and image pyramids are often used to capture objects in different scales. However, image pyramids lead to high computational complexity. For CNN-based methods, brute-force learning (single scale) and image pyramids (multi-scale)

are the most used solutions. Brute-force learning simply forces the network to directly learn scale invariance, which is difficult to learn strong convolutional filters. While multi-scale input always takes a large amount of GPU memories, and spends more time to train and test. Since pedestrians with different scales may show quite different appearances, which lead to varied discriminative power with features extracted from the same descriptor. Especially for small pedestrians, general feature extraction often leads to weak classification. Therefore, how to design a scale adaptive detector is critical for boosting the detection performance.

Among the recent state-of-the-art generic object detectors [13–16], Fast R-CNN [15] and its variant Faster R-CNN [16] are the most prevalent pipelines. One common problem of these solutions is that the last convolutional layer is too small and coarse, which means the features pooled from this layer are lack of sufficient representation capacity for small objects. An intuitive solution is to upsample the feature maps to a proper size. Recently, Long *et al.* [17] proposed an in-network upsampling layer for pixelwise prediction. Noh *et al.* [18] designed a deconvolutional network which contains unpooling and deconvolution operations to decode the convolutional feature maps for generating accurate segmentation results. Similarly to [18], Badrinarayanan *et al.* [19] also proposed an encoder-decoder architecture to achieve pixel-wise segmentation. All these works show that the finer upsampling or decoding , the more accurate predictions. As analyzed above, a refined feature map from the upsampling layers or deconvolutional layers could help to obtain a more accurate portrayal for the small objects. Note that for small objects, the features pooled from a coarse feature map are filled with repeated values which are lack of discriminative representation ability.

Another application is image super-resolution [20], where a coarse to fine deconvolutional layer could capture more rich structural and local information for a finer reconstruction. Therefore, the deconvolutional layers are to obtain a better representation of local and structural information for small pedestrians. It makes sense to design a scale-adaptive network for pedestrian detection. In this way, the features of large and small objects are pooled from different layers for training different regressers respectively, which we called scale-adaptive deconvolutional regression (SADR).

In addition, before the arrival of R-CNN, dollar *et al.* [4] aggregated multiple hand-crafted channels (ACF) to train the cascaded adaboost. Zhang *et al.* [6] applied many of filters in ACF channels to obtain a more rich feature representation. All of these works indicate that the more rich features, the better classification. Recently, feature fusion from different CNN layers has shown the effectiveness to enhance the discriminability[21, 10, 22, 23]. Actually, different layers in a neural network contain different levels of discriminative information. The lower layers always represent the local characteristics, whereas the deeper layers focus on the global semantic information. An intuitive idea is to fuse features of different layers to learn a strong and powerful classifier. Therefore, we investigate the effects of fusing features from different layers on pedestrian classification.

To sum up, the main contributions of this work are as follows:

(1). We propose a novel scale-adaptive deconvolutional regression (SADR) network for pedestrian detection, which could flexibly detect pedestrians with different size.

(2). By computing the classification and regression loss respectively, we integrate multi-layer outputs of CNN network to boost the detection performance.

(3). Extensive experiments on the challenging Caltech dataset well demonstrate the effectiveness of the proposed approach. We achieve the state-of-the-art result with 6.94% miss rate, which is significantly better than 9.29% by CompACT-Deep [24].

2 Related work

In this section, we mainly review some existing object detection and pedestrian detection approaches.

Object detection. Most recent state-of-the-art object detection methods follow the pipeline of R-CNN [13], which first generated object proposals by some unsupervised algorithms (*e.g.* Selective Search [25], Edge Boxes [26] and MCG [27]) from the input image and then classified each proposal into different categories. Since the feature extraction in R-CNN is time-consuming and the training process is implemented through a multi-stage pipeline, two subsequent models, *i.e.* Fast R-CNN [15] and Faster R-CNN [16] were proposed to improve the computational efficiency and integrate the multi-stage training process into an unified pipeline. Moreover, Faster R-CNN introduced a Region Proposal Network (RPN) to reduce the time of proposal generation.

Pedestrian detection. In the literature of pedestrian detection, lots of top performing pedestrian detectors based on hand-crafted features are explored. The Integral Channel Features (ICF) [3] and Aggregated Channel Feature (ACF) [4] efficiently computed features such as local sums, histograms, and Haar features and their various generalizations using integral images. Zhang *et al.* [5] designed informed filters by incorporating prior information as to the appearance of the up-right human body. Cai *et al.* [24] proposed complexity-aware cascaded detectors, which combined features of very different complexities. Deformable part-based models [8] learned a mixture of local templates for each part to deal with appearance variations. Many recent works using convolutional neural networks (CNN) to improve the performance of pedestrian detection [28, 10, 12, 11, 29]. Ouyang *et al.* [12] integrated feature extraction, part deformation handling, occlusion handling and classification into a joint deep model. Tian *et al.* [29] used semantic tasks to assist pedestrian detection. Sermanet *et al.* [10] exploited two contextual regions centered on each object for pedestrian detection. Hosang *et al.* [28] firstly applied the R-CNN framework [13] to pedestrian detection and achieved promising performance on Caltech [30] and KITTI [31] dataset.

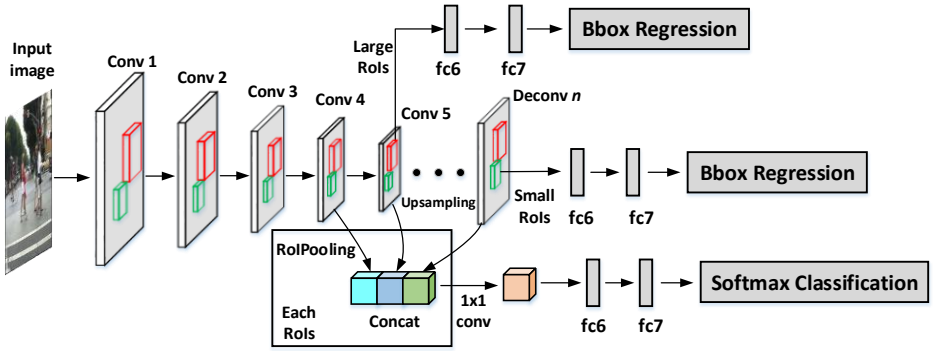

Fig. 1. The architecture of the proposed detection network.

Scale processing. A few works intend to deal with the scale problem. In most cases, image pyramids (multi-scale inputs) are always used to solve this problem, but it is very time-consuming. More recently, Li *et al.* [32] designed two sub-networks to learn the universality and specificity of large-scale and small-scale proposals, which resulted in more training time. Yang *et al.* [33] proposed scale-dependent pooling to handle the scale variation. However, they used the earlier convolutional layers to model the small objects, which might be too weak to make a strong decision. In addition, in order to improve classification performance they also introduced the cascaded AdaBoost Classifiers, which increased the complexity. In this paper, we propose a scale-adaptive deconvolutional regression architecture which is different from [32, 33], and in order to decrease the complexity we just use one classifier.

3 The proposed approach

3.1 Overview

Fig.1 shows the overall framework of the proposed pedestrian detection network. Based on the framework of Faster R-CNN [16], the proposed approach involves two steps: pedestrian candidates generation and pedestrian/background classification. During our implementation, we found that the Region Proposal Network (RPN) serves well as a candidate generator, which achieves 99% recall on Caltech and 96% on KITTI. Thus here we focus on improving the detection accuracy for the second stage. On the one hand, to effectively deal with pedestrians with different scales, we introduce the deconvolutional layers to adaptively upsample the feature map for small pedestrians. In this way, we can adaptively pool RoIs (Region of Intersets) from corresponding layers for regression according to the size of proposals, instead of just pooling from the last convolutional layer. Compared to Fast R-CNN, this scale-adaptive deconvolutional regression (SADR) architecture can more precisely represent pedestrians of different scales and the features used for bounding box regression are more powerful.

On the other hand, to further enhance the discriminative power of classification, we concatenate features RoI-pooled from multiple layers, including deconvolutional and some convolutional layers. This kind of multi-scale feature fusion strategy can effectively capture some fine-grained details by pooling from multiple layers, which is especially important for classification. Conversely, features which are pooled directly from the last convolutional layer may not contain sufficient information for classification.

Finally, the proposed network ends up with three output layers. The first two output layers operate on different RoI feature vectors and output the predicted coordinate tuples for small and large RoIs respectively. The last output layer is the standard softmax layer which predicts the classification score of each RoI.

3.2 Network Architecture

In the proposed detection network, we use the pre-trained VGG16 model [34] to initialize the proposed network. All the convolutional layers and max pooling layers of the VGG16 network are used to encode features before the deconvolutional layers to decode features from the input image. All of the *fc* layers are initialized from VGG16 at the beginning. We randomly initialize the 1×1 convolution layer by drawing weights from a zero-mean Gaussian distribution with standard deviation 0.01. Given an image, we firstly apply RPN to generate a number of proposals. Then the proposed detection network takes the whole image and the proposals as input. In the forward pass, the scale-adaptive deconvolutional regression (SADR) verifies the height of each proposal and pools the region of interest from the corresponding layers according to the height, the details will be discussed in section 3.3. We also concatenate the RoI-pooled features from the outputs of multi-layers to score each proposal.

After the forward pass, each proposal gets a regressed coordinate tuple and a score, which denote the original predictions. Box refinement [35] is followed to the final inference. That is, we take the original predictions as the input and forward into the network again. Thus, we obtain a new classification score and a new regressed box, which denote the new predictions. Then, we merge the new predictions and the original predictions as the final result. Non-maximum suppression (NMS) is applied on the union set of predictions with an IoU threshold of 0.3, and then followed by bounding box voting [36] to refine the final position of proposals. The reason using box voting is that boxes with high scores not always have higher localization accuracy due to various factors, such as training with suboptimal examples and so on. Therefore, it is beneficial for the recall by exploiting predicted boxes around the object to compute the final position.

3.3 Scale-adaptive Deconvolutional Regression

Scale variation, especially for small-scale pedestrians, is a great challenge in pedestrian detection. Focusing on this problem, we propose a novel scale-adaptive deconvolutional regression (SADR) network which effectively integrates small-scale and large-scale regression into a unified framework.

As we known, the max pooling and the stride in convolution operations can reduce the spatial resolution of an input image over layers. As a result, the feature map at the conv5-3 in VGG16 turns out to be 1/16 of the original image. Actually, in Fast R-CNN [15], we can compute the minimal size of object which has no repeated RoI-pooled features using the following equation:

$$S_{min} = s_s \times s_r \tag{1}$$

where s_s and s_r are the stride of specific layer and the output size of RoI pooling operation respectively. The default value of s_r is 7×7. Here we use the height to denote the scale or size of an instance, as the height doesn't vary significantly while the width is more sensitive to pedestrian's pose [30]. Based on the assumption, the minimal height of bounding box is 112 pixels for conv5-3, 56 for the first deconvolutional layer and 28 for the second deconvolutional layer. So the proposed SADR architecture can flexibly choose the appropriate layers to do regression according to the height of pedestrians. The box regression loss is calculated as:

$$L_{loc}(F) = \begin{cases} L_{loc}(F_{c_5}), & h \in [112, \infty) \\ L_{loc}(F_{d_1}), & h \in [56, 112) \\ L_{loc}(F_{d_2}), & h \in (0, 56) \end{cases} \tag{2}$$

where F_{c_5}, F_{d_1} and F_{d_2} denote the features pooled from conv5-3, the 1^{st} and 2^{nd} deconvolutional layers respectively, h means the height of proposals. L_{loc} is the regression loss for each branch. In fact, the Caltech test set (*reasonable*) only evaluates the pedestrians with height greater than 50 pixels, therefore we just use the 1^{st} deconvolutional layer to tackle the instances with height ranging from 50 to 112 pixels.

As shown in Fig.1, each regression branch is followed by 2 4096-d *fc* layers with *ReLU* activations and *dropout* layers so as to learn a set of scale-specific parameters. In the fine-tuning process, we first divide the input proposals into three groups depending on the height and then feed them into corresponding RoI pooling layers so as to pool features from corresponding layers for regression. Since each regression branch learns from scale-specific samples, the branch will capture the rich information for this scale, resulting in a more accurate detection model.

The advantages of SADR are two-fold. On the one hand, the features extracted from deconvolutional layers could capture more rich structural information which are more powerful than the earlier convolutional layers. On the other hand, a single network is designed to adaptively select corresponding layers for regression according to the height of RoIs, which makes the training process more concise and fast. In this way, the proposed SADR effectively avoids upsampling the input images to handle specific instances.

3.4 Multi-layer Feature Fusion

As analysis previously, using reduplicative RoI-pooled features is hard to learn a strong classifier and regresser. We concatenate the RoI-pooled features from

multiple layers, including deconvolutional and some convolutional layers. This kind of multi-scale feature fusion strategy can effectively capture some fine-grained details by pooling from multiple layers, which are especially important for classification. These multi-layer features are used to predict its scores to pedestrian.

The implementation of fusing multi-layer features is closely related to those used in [21]. Different from [21], we find it also works well without the complex operations of L2-normalized and scaled. That's to say, we directly concatenate each RoI-pooled feature along the channel axis and reduce the dimension to $512 \times 7 \times 7$ with a 1×1 convolution. This kind of integration for multi-layer features can effectively boost the performance of pedestrian detection.

3.5 Multi-task Loss

For training the proposed detection network, we use a multi-task loss for joint object classification and scale-adaptive bounding box regression. The final loss function is defined as:

$$L(p, k^*, t, t^*) = L_{cls}(p, k^*) + \sum_{i=1}^{n} \lambda_i [k_i^* \geq 1] L_{loc\_i}(t_i, t_i^*) \tag{3}$$

where $p = (p_0, p_1)$ is a discrete probability distribution over 2 categories (pedestrian or not). k^* is the true category label, n is the number of scales, $k_i^* \subseteq k^*$ is the true label of RoIs corresponding to i^{th} scale. t_i and t_i^* are the predicted tuple and the true tuple for bounding box regression respectively. L_{loc_i} is the standard smoothed L_1 loss. λ_i is the predefined hyper-paremeter to balance the losses of different tasks, here we set $\lambda_1 = \cdots = \lambda_n = 1$ as same to Fast R-CNN.

It should be noted that the proposed network achieves regression and classification in a different manner. Only the regression distinguishs the size of objects, while the classification loss treats all objects equally by exploiting identical features regardless of the size of objects. That's because classification and regression are two different sub-tasks. So in backpropagation, derivatives for classification loss can be back-propagated to all the previous layers.

4 Experiments

4.1 Datasets and Metrics

Caltech pedestrian dataset. The Caltech dataset [30] is one of the most prevalent datasets for pedestrian detection. It consists of 10 hours of 640x480 30Hz video in an urban traffic environment. The raw annotations amount to a total of 350k bounding boxes and 2300 unique pedestrians. The standard training set and test set extract one out of each 30 frames, which results in 4024 frames with 1014 pedestrians for evaluating. In most case, researchers can leverage more data for training by extracting one out of three or four frames. Recently, Zhang *et at.* [1] revised the original annotations and released a new high quality ground

truth for training and testing. In this paper, we use the new annotations of Caltech10x for training and evaluated on the new aligned test set. In the standard Caltech evaluation, the log-average miss rate (MR) which is averaged over the FPPI range of $[10^{-2}, 10^{0}]$ is used to evaluate the performance of detectors.

KITTI dataset. The KITTI object detection benchmark [31] consists of 7481 training images and 7518 test images. Due to the diversity of scale, occlusion and truncation of objects, the dataset evaluates at three levels of difficulty, *i.e.*, easy, moderate and hard, where the difficulty is differentiated by the minimal scale of object and the occlusion and truncation of the object. The benchmark follows the PASCAL protocol and use Average Precision (AP) to measure the detection performance, where 50% overlap thresholds are adopted for pedestrian.

4.2 Evaluation on Caltech Dataset

Implementation details. We use the pre-trained VGG16 model to initialize the proposed detection framework. For the RPN stage, each location in a feature map generates 10 bounding boxes with one aspect ratio of 2.0, where 2.0 indicates the ratio between height and width of the box. In both RPN and the proposed detection network, the scale of input image is set to be 720 pixels on the shortest side and the negative examples have an IoU threshold ranges from 0 to 0.1. We use stochastic gradient descent with momentum of 0.9 and weight decay of 0.0005 to train our detection network. Each SGD mini-batch is constructed from 2 images which are randomly chosen from the whole training set. The foreground-to-background in a mini-batch is set to be $1:3$, thus ensuring that 25% of training examples is foreground (*fg*) RoIs.

For training detection network, we use the same way as [16]. In stage 1, we update all the paremeters except the first four convolutional layers. We fine-tune the network for about 4 epochs with initial learning rate of 0.001 which is decayed by 0.1 after 2 epochs. In stage 2, we only update the parameters of deconvolution and *fc* layers. We fine-tune the network for about 10 epoches with a fixed learning of 0.0001. The whole network is trained on a single NVIDIA GeForce GTX TITAN X GPU with 12GB memory.

Analysis on SADR. We first evaluate the effectiveness of the proposed SADR architecture. Table 1 shows that the Faster R-CNN baseline [16] which just uses the feature map of conv5-3 to do classification and regression achieves 10.59% miss rate, which outperforms most of state-of-the-art detectors showed in Fig. 3. We observe that the main factor affecting performance is small objects, since the large objects have already achieved 1.10% miss rate. Therefore we upsample the last convolutional feature map by inserting a deconvolutional layer between conv5-3 and fc6 in VGG16. The features pooled from the 1^{st} deconvolutional layer are fed into the classifier and regresser. As we predicted, the overall performance improves a lot. Especially for the small objects, the miss rate improves from 11.65% to 10.29%, which verifies that a bigger feature map is better for

locating small objects. We also observe that the performance of large objects degrades slightly. It maybe lost some spatial information when pooling from the deconvolutional layer, since the last convolutional layer can normally pool object with minimum height of 112 pixels. With a flexible scale-adaptive strategy, the best performance is achieved by the proposed SADR method both in small and large objects. On the one hand, compared with the baseline Faster R-CNN, the performance of large objects is further improved and achieves zero miss rate, which implies that the features pooled from the deconvolutional layer are more discriminative for classification than the last convolutional layer. It indirectly indicates that features extracted from deeper structures are more powerful to differentiate the object category. On the other hand, in contrast to the second model in Table 1, the performance of small objects is further improved from 10.29% to 9.04%, which confirms that the SADR architecture can more easily learn the inter-scale variances as the small-scale and large-scale branches just focus on their own specificity respectively.

Table 1. Miss rate(MR) of baseline and our scale-adaptive deconvolution regression(SADR) based model. S, L and A denote the height group of $[50, 112)$, $[112, \infty)$ and $[50, \infty)$. baseline: Faster R-CNN [16]; d_1: the 1^{st} deconvolutional layer.

Methods	SADR	S	L	A
baseline[16]	no	11.65%	1.10%	10.59%
baseline[16]+d_1	no	10.29%	1.55%	9.41%
baseline[16]+d_1	yes	**9.04**%	**0.0**%	**8.27**%

Analysis on Multi-layer Feature Fusion. We also compare different feature fusion strategies for the classification performance. In this section, we use the same SADR architecture, just to verify the effectiveness of multi-layer feature fusion for classification. Based on the analysis of Table 1, we use features pooled from conv5-3 to do bounding box regression for candidates with height greater than 112 pixels and use the first deconvolutional layer for height ranging from 50 to 112 pixels.

As shown in Table 2, the first line is the baseline Faster R-CNN, the rest of models all use the proposed SADR architecture. As the analysis previously, the model in line 2 outperforms baseline for two reasons. First, features extracted from the finer feature map are more discriminative for classification. Second, the SADR architecture is superior in learning the inter-scale variances. We observe that the performance almost remains unchanged by blending Deconv1 with Conv5, while the fusion of Deconv1 and Conv4 benefits a little more. This result is caused by two factors, the layer depth and down-sampling factor. In general, a deeper layer with lower down-sampling factor can better represent the object,

there is a tradeoff between depth and down-sampling factor. This observation also stands in CCF [37]. The fusion of these three layers achieves the best performance, which confirms that the more rich features, the better classification. In fact, different layers in convolutional neural networks contain different levels of structural information, while integrating them can capture fine-grained details.

Box refinement. As shown in Table 2, we observe that MR improves a lot and reaches 8.96% after adding box refinement for baseline. We all know that box refinement just simply revises the position of final detection boxes, it has no effects on the score of bounding box. In addition, combined with Table 1, we can conclude that the main factor influencing the performance of a detector is the regression part of the network, *i.e.*, the ability of localization for small objects. However, for model in line 2, the performance degrades slightly after adding box refinement, which means the accuracy of localization is good enough. Fig. 2 shows the average miss rate for different feature fusion strategies.

Table 2. Comparison performance for different feature fusion strategies. We only use two scale-adaptive regression branches. Line 1 is the baseline Faster R-CNN; line 2 to line 5, using Conv5 to do regression for height between $[112, \infty)$ and Deconv1 for height less than 112. Deconv1: features pooled from the first deconvolution; Conv5: features pooled from conv5-3 in VGG16; Conv4: features pooled from conv4-3 in VGG16; MR: log-averaged miss rate over the FPPI range of $[10^{-2}, 10^{0}]$; BR: box refinement, first introduced in [36] as iterative localization.

Deconv1	Conv5	Conv4	MR	+BR	Δ
	√		10.59%	8.96%	+1.63%
√			8.27%	8.48%	−0.21%
√	√		8.25%	8.48%	−0.23%
√		√	7.96%	7.90%	+0.06%
√	√	√	**7.40**%	**6.94**%	+0.46%

Comparison with the State-of-the-arts. The overall experimental results are shown in Fig. 3. We compare our detector with all the existing best-performing methods, including hand-crafted models, like ACF-Caltech+ [38], LDCF [38], Katamari [39], SpatialPooling+ [40](which combines HOG, LBP, spatial covariance and optical flow) and Checkboards [6](which requires a large number of filter channels), and CNN-based models, like TA-CNN [29] and the current state-of-the-art CompACT-Deep [24]. Our method outperforms the current best detector CompACT-Deep by 1.89%, with the help of box refinement strategy, we further lower the MR to 6.94%. Since we use the new revised Caltech10x to train our model, we just evaluate on the new testing set.

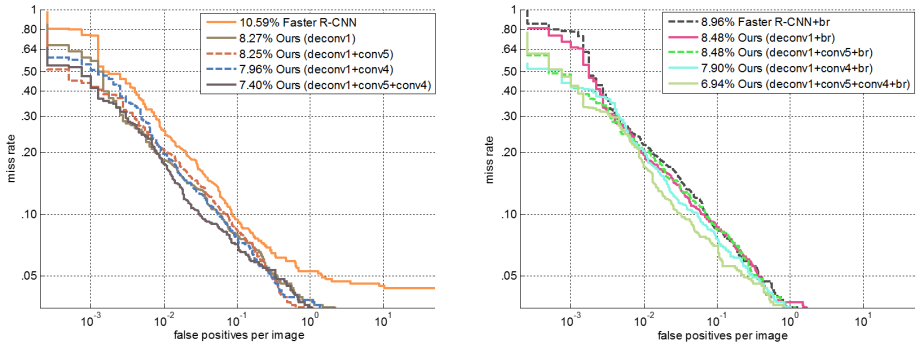

Fig. 2. Average miss rate on Caltech test set. (a) the comparison of different feature fusion strategies; (b) adding box refinement.

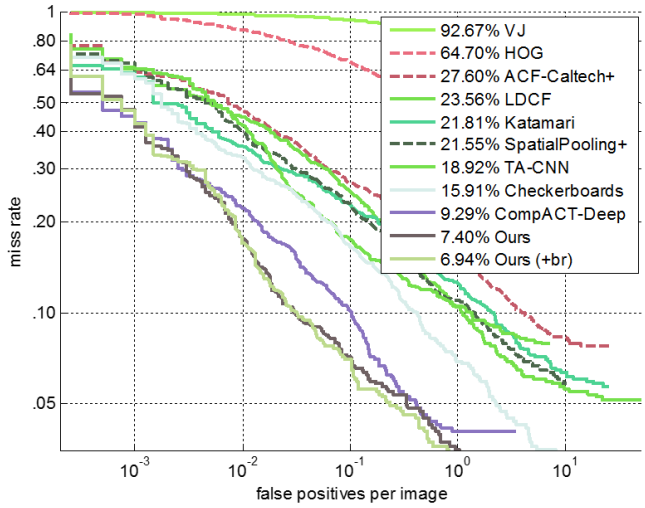

Fig. 3. The comparison of pedestrian detection performance with all recent state-of-the-art methods on revised Caltech test set [1].

4.3 Evaluation on KITTI Dataset

We also evaluate our method on the more challenging KITTI dataset [31]. Since KITTI contains more small objects with minimum height 25 pixels, we insert two deconvolutional layers in our network to ensure the features pooled from corresponding layers are more powerful. Therefore, we group the object proposals into 3 levels based on their height, *i.e.*, $[25, 56)$ for the 2^{nd} deconvolutional layer, $[56, 112)$ for the 1^{st} deconvolutional layer and $[112, \infty)$ for conv5-3. The scale

of input image is set to be 450 pixels on the shortest side, other parameter configurations are the same as Caltech.

Table 3 and Fig. 4 show the performance on KITTI. It can be observed that our method achieves a competitive result, which outperforms the SDP+RPN [33] on Easy and Moderate subsets. In contrast with the CompACT-Deep [24], which is the state-of-the-art on Caltech, our methods improves 12.94%, 11.96% and 11.96% on Easy, Moderate and Hard subsets respectively. Our detector on KITTI consists of 3 scale-specific regression branches, which means more parameters need to be learn than Caltech. However, KIITI has less training data which only contains 7481 training images covering 4487 pedestrians, our network can achieve a better result with more training images.

Table 3. Comparison to state-of-the-art on KITTI Pedestrian. The evaluation metric is average presion (AP). Note: * indicates anonymous submission and this paper is submitted to *ECCV16.*

Methods	Easy(%)	Moderate(%)	Hard(%)
Ours	83.63	70.70	64.67
SDP+RPN[33]	80.09	70.16	64.82
3DOP[41]	81.78	67.47	64.70
Mono3D[42]	80.35	66.68	63.44
Faster R-CNN*[16]	78.86	65.90	61.18
SDP+CRC(ft)[33]	77.74	64.19	59.27
CompACT-Deep[24]	70.69	58.74	52.71
FilteredICF[6]	67.65	56.75	51.12

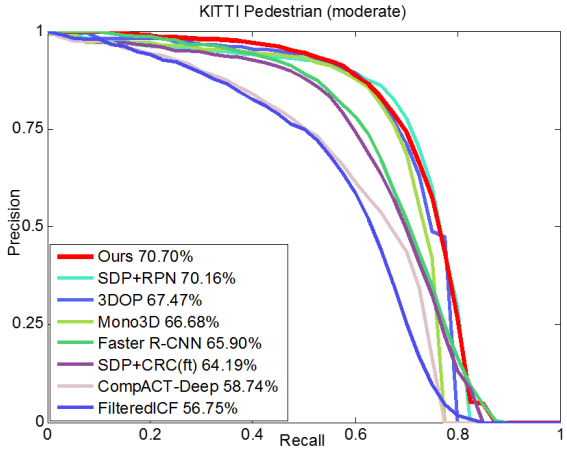

Fig. 4. Comparison to state-of-the-art on KITTI pedestrian (moderate).

5 Conclusion

In this paper, we propose a novel scale-adaptive deconvolutional regression (SADR) network for pedestrian detection, which could flexibly detect pedestrians with different size. Since each regression branch learns from scale-specific examples, the proposed network has the ability to capture inter-scale differences, resulting in a more sophisticated detection model. By computing the classification and regression loss respectively, we integrate features pooled from multi-layers to further boost the detection performance. Extensive experiments on the public pedestrian dataset clearly demonstrate the superiority of the proposed method, we achieve a state-of-the-art result with MR 6.94% on Caltech dataset and a promising result on KITTI.

6 Acknowledgment

This work was supported by 863 Program 2014AA015104, and National Science Foundation of China 61273034, and 61332016.

References

1. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: CVPR. (2016)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 1., IEEE (2005) 886–893
3. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. (2009)
4. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on **36** (2014) 1532–1545
5. Zhang, S., Bauckhage, C., Cremers, A.: Informed haar-like features improve pedestrian detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 947–954
6. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE (2015) 1751–1760
7. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
8. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32** (2010) 1627–1645
9. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Cascade object detection with deformable part models. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, IEEE (2010) 2241–2248
10. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 3626–3633

11. Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 3258–3265
12. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 2056–2063
13. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 580–587
14. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **37** (2015) 1904–1916
15. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1440–1448
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. (2015) 91–99
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3431–3440
18. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Computer Vision (ICCV), 2015 IEEE International Conference on. (2015)
19. Badrinarayanan, V., Handa, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293 (2015)
20. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. (2015)
21. Bell, S., Zitnick, C.L., Bala, K., Girshick, R.: Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. arXiv preprint arXiv:1512.04143 (2015)
22. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 447–456
23. Zagoruyko, S., Lerer, A., Lin, T.Y., Pinheiro, P.O., Gross, S., Chintala, S., Dollár, P.: A multipath network for object detection. arXiv preprint arXiv:1604.02135 (2016)
24. Cai, Z., Saberian, M., Vasconcelos, N.: Learning complexity-aware cascades for deep pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 3361–3369
25. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision **104** (2013) 154–171
26. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: Computer Vision–ECCV 2014. Springer (2014) 391–405
27. Arbeláez, P., Pont-Tuset, J., Barron, J., Marques, F., Malik, J.: Multiscale combinatorial grouping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 328–335
28. Hosang, J., Benenson, R., Omran, M., Schiele, B.: Taking a deeper look at pedestrians. In: CVPR. (2015)
29. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 5079–5087

30. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34** (2012) 743–761
31. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 3354–3361
32. Li, J., Liang, X., Shen, S., Xu, T., Yan, S.: Scale-aware fast r-cnn for pedestrian detection. arXiv preprint arXiv:1510.08160 (2015)
33. Yang, F., Choi, W., Lin, Y.: Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. (2016)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
35. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
36. Gidaris, S., Komodakis, N.: Object detection via a multi-region and semantic segmentation-aware cnn model. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1134–1142
37. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Convolutional channel features. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 82–90
38. Nam, W., Dollár, P., Han, J.H.: Local decorrelation for improved pedestrian detection. In: Advances in Neural Information Processing Systems. (2014) 424–432
39. Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? In: Computer Vision-ECCV 2014 Workshops, Springer (2014) 613–627
40. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Strengthening the effectiveness of pedestrian detection with spatially pooled features. In: Computer Vision–ECCV 2014. Springer (2014) 546–561
41. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals for accurate object class detection. In: NIPS. (2015)
42. Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., Urtasun, R.: Monocular 3d object detection for autonomous driving. In: CVPR. (2016)