Occluded Pedestrian Detection Through Guided Attention in CNNs

Shanshan Zhang[1], Jian Yang[2], Bernt Schiele[3]
[1]Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education,
[2]Jiangsu Key Lab of Image and Video Understanding for Social Security,
School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
shanshan.zhang@njust.edu.cn, csjyang@njust.edu.cn, schiele@mpi-inf.mpg.de

Abstract

Pedestrian detection has progressed significantly in the last years. However, occluded people are notoriously hard to detect, as their appearance varies substantially depending on a wide range of occlusion patterns. In this paper, we aim to propose a simple and compact method based on the FasterRCNN architecture for occluded pedestrian detection.

We start with interpreting CNN channel features of a pedestrian detector, and we find that different channels activate responses for different body parts respectively. These findings motivate us to employ an attention mechanism across channels to represent various occlusion patterns in one single model, as each occlusion pattern can be formulated as some specific combination of body parts. Therefore, an attention network with self or external guidances is proposed as an add-on to the baseline FasterRCNN detector. When evaluating on the heavy occlusion subset, we achieve a significant improvement of 8pp to the baseline FasterRCNN detector on CityPersons and on Caltech we outperform the state-of-the-art method by 4pp.

1. Introduction

Pedestrian detection has been attracting intensive interests in both academia and industry. During the last decade, great progress has been achieved [3, 31], especially by properly adapting CNNs for general object detection to this canonical task [12, 30, 32]. Although the state-of-the-art performance is plausible across different datasets, we observe that it drops significantly as occlusion grows. In real world applications, occlusion happens very often but is challenging to handle. To encourage more work for occlusion handling, the CityPersons dataset has been proposed [32], which consists of a large number of occlusion cases with various patterns.

Some efforts have been made to handle occlusion, but most of them train ensemble models for most frequent occlusion patterns [7, 19, 21, 26]. The major drawback of those methods is that it is very time-consuming at both training and testing times. Some other works propose to model different occlusion patterns in a joint framework [33, 22], but they still rely on an integration of a set of occlusion/part detection scores. The above methods are not able to cover all occlusion patterns, and the independent integration procedure inhibits error propagation to the occlusion/part detection module.

Figure 1. Visualization of detection results (at FPPI=0.1) from our FasterRCNN+ATT-part detector (upper row) and the baseline FasterRCNN detector (lower row). Our attention based detector achieves higher recall for occluded pedestrians. The sample images are selected from the CityPersons [32] validation set; we show ground truth annotations in green and detection results in red. All figures are best viewed in color.

In order to deal with a wide range of frequent and less frequent occlusion patterns in one coherent model, we propose different attention mechanisms, which guide the detector to pay more attention to the visible body parts. These attention mechanisms are motivated by the fact that different channels of CNN-based detectors, in our case a FasterRCNN pedestrian detector, are selective and show strong responses for different body parts. These explicit representations of body parts motivate us to propose channel-wise

attention mechanisms to learn proper attention parameters for different channels so as to handle different occlusion patterns effectively.

1.1. Contributions

In summary, our contributions are as follows:

(1) We provide an analysis to understand the relation between body regions and different CNN channel features of a pedestrian detector, and find many of them are localizable and interpretable.

(2) We apply channel-wise attention mechanism to handle different occlusion patterns by adding an additional attention net to the FasterRCNN architecture. We explore different attention guidances, including self attention, visible-box attention and part attention. Our method only makes minor changes to the vanilla FasterRCNN architecture, thus is easy to implement and train.

(3) We achieve state-of-the-art results on several standard benchmarks; the improvement is significant especially on the heavy occlusion subset. On CityPersons, the proposed method achieves 8pp gain compared to the FasterRCNN baseline; on Caltech, we outperform the state-of-the-art method by 4pp.

To the best of our knowledge, this paper is the first attempt to analyze channel-wise attention for pedestrian detection, and it is the first work to handle occlusion in the FasterRCNN architecture.

1.2. Related Work

Since we use a FasterRCNN detector as our base pedestrian detector, and an attention network as an add-on to handle occlusion, we review recent work on CNN based pedestrian detectors, occlusion handling for pedestrians and attention mechanisms respectively.

Pedestrian detection with CNNs. Convolutional neural networks (convnets) have achieved great success in the general object detection task on the ImageNet [17], Pascal, and MS COCO datasets [11]. Early works [12, 30] applying convnets for pedestrian detection are based on the RCNN structure [9], which relies on high-quality external proposals to achieve good performance. More recently, FasterRCNN [23] has become the de-facto standard architecture, which allows end-to-end learning. At the meantime, some works present good results using customized architectures, such as MS-CNN [4] and SA-FastRCNN [18]. However, proper modifications to the vanilla FasterRCNN [32] allow to reach state-of-the-art results in pedestrian detection. Therefore, we follow [32] and use the adapted FasterRCNN architecture for experiments in this paper.

Occlusion handling for pedestrian detection. The most common occlusion handling strategy is to learn a set of detectors, each corresponding to one specific manually designed occlusion pattern. Different features are employed, including hand-crafted features [7, 19] and deep convolutional features [21, 26]. The final decision is made by integrating the output of these ensemble models. The drawback of those methods is that each part/occlusion pattern detector is learned independently, and it is time consuming to apply the set of models at test time. On the other hand, some other works proposed to learn multiple occlusion patterns in a joint way [33, 22], which saves a lot of training and testing time. However, the final decision is still made by integrating multiple part scores, which makes the whole procedure more complex and hard to train. In contrast we learn a continuous attention vector that is both easy to train and also has low overhead.

Attention mechanisms in CNNs. The attention mechanism has been widely used in CNNs for different computer vision tasks, for instance, object detection [2], digits recognition [15], and pose estimation [20]. The above works all investigate to model spatial correlations. In contrast, [13] proposes squeeze-and-excitation networks to model the interdependencies between channels of convolutional features. However, the channel-wise attention is self guided, i.e. no external signal is employed. In contrast, in this paper we will show that external guidance can help to improve the performance of channel-wise attention mechanisms.

2. Body Parts and Channel Features

Convnets have shown to be capable of learning representative features for object detection, and some recent works analyze the interpretability of the hidden neurons by visualizing their activations. For instance, [1] performs network dissection and finds that many individual units respond to specific high-level concepts and [28, 10] also find some filter responses can be linked to semantic parts.

Similarly, in this paper, we investigate whether channels can be related to human body parts in a pedestrian detector. We first train a FasterRCNN (VGG16) detector on the CityPersons training set. After training, we pick one arbitrary image from the CityPersons validation set, which contains multiple people, and let it pass the network for feature extraction. As default, on the top convolutional layer, we have 512 channels in total.

In the following, we examine the activations of each channel respectively. As shown in figure 2 for three representative channels, the original image is overlaid with the activation map. From the visualizations, we make the following observation: Many channels show some highly localizable activation pattern, relating them to specific body regions or body parts; the three channels show strong activations at people's head, upper body and feet respectively. Similar findings are shown in [25], that in a bird classification network some channels are associated with parts.

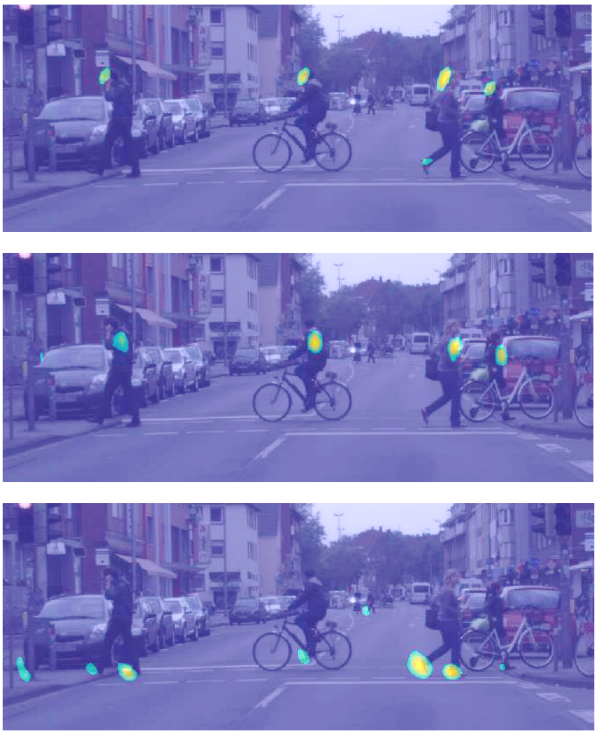


Figure 2. Relation to body parts of different channel features from a FasterRCNN pedestrian detector. Highlighted regions trigger strong activation inside each channel.

To better understand the relation between body parts and all channels in a statistical way, we implement pixel-wise XOR operation between each binarized channel feature map and part detection heatmap [14]. The correlation value for each pair is measured by the percentage of one values in the XOR map. We find for each image, more than 30% channels show strong correlation (correlation value$\geq 60\%$) with one of 14 part detection heatmaps.

This observation encourages us to explore the possibility of channel-wise attention for occluded pedestrian detection as such an attention mechanism that can focus more on the visible body regions and focus less on the occluded regions.

3. Guided Attention in CNNs for Occlusion Handling

The major challenge of handling occlusion comes from the large variety of occlusion patterns, which lead to rather diverse appearances of human bodies, as shown in figure 3. In this paper, we propose to employ channel-wise attention in convnets allowing the network to learn more representative features for different occlusion patterns in one coherent model.

3.1. Overview

The FasterRCNN detector obtains state-of-the-art results in pedestrian detection [32]. In this paper, we use it as the base detector in our experiments, while adding an attention network as a separate component to generate a channel-wise attention vector. The flowchart of our FasterRCNN detector with an attention network is shown in figure 4. The upper flow is a typical feature extraction procedure of a FasterRCNN detector: first, the input images go through the base net (e.g. VGG16); and then a region proposal network (RPN) is used to generate proposals; after that, the features for each proposal are generated by cropping from the top convolutional feature maps and a following RoiPooling layer produces the same length of features for each proposal. These features will go through the classification network for category prediction and bounding box regression. The FasterRCNN network can be trained end-to-end by optimizing the following loss function:

(a) (b) (c)

Figure 3. Different occlusion patterns lead to variation of human appearance.

$$L_0 = L_{rpn\_cls} + L_{rpn\_reg} + L_{cls} + L_{reg}, \quad (1)$$

where: L_{rpn_cls} and L_{cls} are the cross-entropy loss for classification in the RPN and main network; L_{rpn_reg} and L_{reg} are the L_1 loss for bounding box regression.

In our methods, an additional attention net is proposed to regress the channel-wise attention vector, namely Ω, which is used to apply a re-weighting operation on the multi-channel convolutional features. After the re-weighting procedure, the features are passed to the classification network.

3.2. Channel-Wise Attention

As discussed in section 2, many channels in a pedestrian CNN are localizable and can be related to different body parts. This observation strongly motivates us to perform re-weighting of channel features to handle various occlusion patterns. Let occlusion pattern n be defined as the following vector:

$$occl(n) = [v_0 p_0, v_1 p_1, ..., v_k p_k], v_i \in \{0, 1\}, i \in [0, k], \quad (2)$$

where p_i represents each body part and v_i is a binary variable, indicating the visibility of the ith part.

In typical CNNs, channels' weights are fixed and thus do not vary across different samples. This mechanism limits

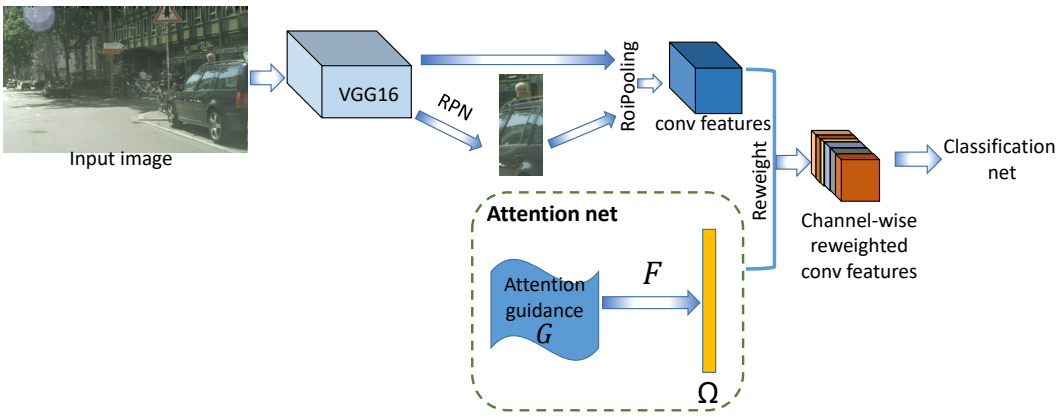
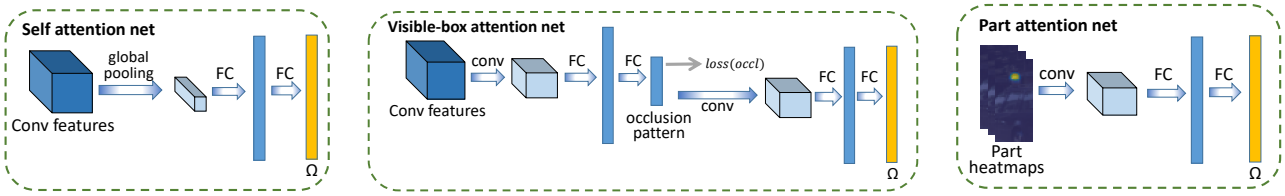


Figure 4. Flowchart of attention guided FasterRCNN pedestrian detector. An attention network is added to the FasterRCNN architecture to generate the weighting parameters Ω for the top conv features.



(a) (b) (c)

Figure 5. Three different attention nets use different attention guidances.

the network's adaptivity to various appearances. For example, when the person's body is occluded as in figure 3(a), the feet channel contributes to the final score irrespective of the occlusion. This, however, will typically result in a lower overall score as the occlusion patterns are too variable to allow to generate an equally high score as for un-occluded pedestrians.

Our intuition is to allow the network to decide for each sample, how much each channel should contribute in the final feature pool. Intuitively, the network should let those channels representing the visible parts contribute more, while the invisible parts contribute less.

The re-weighting of channels can be presented as follows:

$$f_{occl}(n) = \Omega_n^T f_{chn}, \tag{3}$$

where f_{chn} indicates the top channel features, and Ω_n is the weighting parameter vector for the nth occlusion pattern.

In this way, the importance of the channel features varies for each sample as its occlusion pattern changes. For example, when the left body is occluded, Ω should be adjusted so that the corresponding channels representing the left body region have lower weights, which means they have lower impact on the final score.

3.3. Attention Networks

The attention network is an important component in our method to generate the attention vector Ω. As shown in the lower part of figure 4, the attention network takes an input of attention guidance G, and then learns a mapping function F used to regress Ω as output:

$$\Omega = F(G^T). \tag{4}$$

While we have motivated the attention vector Ω being related to specific occlusion patterns, it is important to note, that our attention vector Ω in all our attention networks is continuous and thus not restricted to any particular discrete set of occlusion patterns as some previous work [26, 33, 22]. Instead, the attention vector Ω is trained end-to-end for all our attention networks either through self-attention or guided by some additional external information.

We consider three different types of guidance G: (1) top convolutional features; (2) visible bounding boxes; (3) part detection heatmaps. Depending on which information we use as guidance, we define our attention nets as: self attention, visible-box attention and part attention nets, respectively. We start with self attention, and then further exploit to use external information as stronger guidance. We show an illustration for the above three attention nets in figure 5.

3.3.1 Self Attention Net

SENet is the first attempt to exploit channel-wise attention in CNNs [13]. The goal is to enhance the representational ability for various samples by explicitly modelling the interdependencies between the convolutional channels. To this end, a "Squeeze-and-Excitation" (SE) block is proposed to perform sample-dependent feature re-weighting, through which the more informative features are selected while less useful ones are suppressed. The SE block is composed of one global average pooling layer and two consecutive fully connected layers. SENet is easy to implement, obtaining remarkable improvements while adding little additional computational costs.

Inspired by SENet, we design our self attention net to learn the channel-wise attention parameters Ω. It is a re-implementation of SENet with an identical block structure. Since no external information is needed, we call it self attention. We show the self attention net in figure 5(a), where we use conv5_3 features as guidance G to regress Ω.

We refer to the FasterRCNN detector using the self attention net as FasterRCNN+ATT-self in this paper.

3.3.2 Visible-box Attention Net

The self attention net models the channel-wise attention using the channels themselves, while we believe the attention network's capacity can be improved with external information as additional input or supervision. Intuitively, one useful guidance to regress Ω should be the occlusion patterns themselves, as they contain information about visibility of body parts. Ideally, occlusion patterns should be defined as in equation 2, by indicating the visibility of each body part. However, in practice it is too expensive to obtain body part annotations. Alternatively, we define it coarsely by the combination of one full body bounding box along with one visible box, which are provided in some popular pedestrian datasets. Since we use the visible box as external guidance, we refer to this net as visible-box attention net.

However, the visible box is not available at test time, thus the occlusion pattern can not be simply used as input to the attention net. To overcome this problem, we propose to learn the occlusion pattern in a supervised manner inside the attention net. By analyzing the training data on the CityPersons dataset, we find the most frequent occlusions are as follows: (1) fully visible; (2) upper body visible; (3) left body visible; (4) right body visible. The other patterns are ignored as too little training data is available. In this way, the occlusion pattern estimation is formulated as a four-class classification task.

The visible-box attention network architecture is shown in figure 5(b), where the occlusion pattern estimation subnet consists of one convolutional and two fully connected layers. Once the occlusion pattern is estimated, one convolutional layer is used for feature extraction followed by two fully connected layers to regress Ω. In this way, we add one more task of occlusion estimation to the pipeline, and the loss function of the whole system can be written as follows:

$$L_{ATT-vbb} = L_0 + \alpha L_{occl}, \tag{5}$$

where L_0 is the loss function used in vanilla FasterRCNN (equation 1), and L_{occl} is defined as cross-entropy loss for occlusion pattern classification. All the parameters in the network are optimized in an end-to-end fashion. We set $\alpha = 1$ by default.

We refer to the FasterRCNN detector using visible-box attention net FasterRCNN+ATT-vbb in this paper. It is worth noting that, as a side-effect, an estimate of the occlusion pattern is obtained at test time, not provided by previous methods.

3.3.3 Part Attention Net

Making use of visible bounding boxes allows us to train an occlusion pattern estimation subnet, which serves as a guidance to regress the continuous attention vector Ω. However, there are two problems with visible bounding boxes: (1) It is expensive to obtain visible boxes as additional training annotations; (2) Sometimes occlusion happens irregularly, resulting in that the visible part can hardly be covered by one single rectangular box, see two examples in figure 6.

To overcome the above two problems, we investigate to estimate the occlusion pattern by using body part detection results, which are supposed to predict the visibility of each body part, e.g. head, shoulder, arm, etc.

In principle, we can implement our part attention net in the same way as the visible-box attention net, inside which the occlusion patterns can be estimated and immediately used as guidance to regress Ω. However, on pedestrian detection datasets, we do not have body part annotations for supervision, so we decide to use a pre-trained part detection network trained on the MPII Pose Dataset [14]. This detector is a fully convolutional network, providing precise predictions for 14 human body keypoints. We apply this part detector without any changes or finetuning on the CityPersons dataset and achieve surprisingly good results. We show two examples in figure 6.

From figure 6, we can see that the two persons occluded by a pole and a car still trigger rather strong response at the location of the visible parts on the heatmaps. These results inform us that when a full body detector fails for an occluded person, the part detector is still able to make precise predictions for visible parts. Therefore, the part detection heatmaps can be used as an effective hint of occlusion patterns to guide the attention network.

The attention network using part detections is shown in figure 5(c), where 14 keypoint heatmaps are used as input.

Figure 6. Occluded persons show strong response on heatmaps for visible parts. We use pretrained part detectors from the DeeperCut paper [14]. For some cases, the visible part cannot be covered by one single rectangle, while part heatmaps represent the occlusion patterns more precisely.

As we assume that the spatial information plays an important role for guidance we apply one convolutional layer for feature extraction and two fully connected layers to regress the continuous attention vector Ω. This is in contrast to the self attention net that uses global pooling instead.

We refer to the FasterRCNN detector using part attention net FasterRCNN+ATT-part in this paper.

4. Experiments

In this section, we will first introduce the evaluation metrics we use, followed by a brief description to the datasets used for experiments, and some implementation details. After that, we will show experimental results for the different attention networks, and make a comparison to the state of the art. In the end, we will visualize how attention works in our detectors.

4.1. Evaluation Metrics

We use the standard average-log miss rate (MR) in all of our experiments, which is computed in the FPPI range of $[10^{-2}, 10^{0}]$ [6]. Since we care more about occluded pedestrians in this paper, we will show our results across different occlusion levels:

(1) Reasonable (**R**): visibility $\in [0.65, inf]$;

(2) Heavy occlusion (**HO**): visibility $\in [0.20, 0.65]$;

(3) Reasonable+Heavy occlusion (**R+HO**): visibility $\in [0.20, inf]$.

The performance on the **R+HO** subset is used to measure the overall performance as it includes a wide range of occlusions. Note that we only consider pedestrians with height $\in [50, inf]$ for all experiments.

4.2. Datasets

CityPersons. We use the CityPersons dataset [32] for most of our experiments. The CityPersons dataset was built upon the Cityscapes dataset [5], which was recorded in multiple cities and countries across Europe and thus shows high diversity. Importantly, it includes a large number of occlusion cases. We use the original training and validation split, which are composed of 2,975 and 500 images respectively.

	# images	**R**	**HO**	**R+HO**
CityPersons	500	1579	733	2312
Caltech	4024	1014	273	1287
ETH	1804	-	-	11941

Table 1. Comparison of the number of pedestrian boxes in each evaluation subset on the CityPersons validation set, Caltech test set and ETH dataset. Note, that for ETH the numbers on R and HO subsets are not provided as visible boxes are not available.

Caltech. The Caltech [6] dataset is one of the most popular ones for pedestrian detection. It consists of approximately 10 hours of 640×480 30Hz video taken from a vehicle driving through Los Angeles. We use set00-set05 for training and sample with 10Hz to get a large amount of training data (42,782 images in total). The test set consists of 4,024 images sampled with 1Hz from set06-set10.

ETH. The ETH dataset [8] "Setup 1 (chariot Mk I)" consists of three sequences (1804 images in total) for testing. As the images were captured in the city center, it contains intensive crowds, thus a suitable test base for occluded pedestrian detection.

In table 1, we show statistics on the evaluation subsets for different datasets. Although Caltech provides more images for testing, the number of occlusion cases is smaller than that on the CityPersons dataset.

4.3. Implementation Details

On the CityPersons dataset, we finetune from the ImageNet model with the Adam solver [16]. We train with an initial learning rate of 10^{-3} for $20,000$ iterations and train for another $5,000$ iterations with a decreased learning rate of 10^{-4}; we do not upsample the input images, as it is more than 2x faster for both training and testing, resulting in only a small performance drop of $\sim$1pp.

On Caltech, we finetune from the CityPersons model. We start with a small learning rate of 10^{-4}, and then decrease the learning rate after every $20,000$ iterations. The model converges at $30,000$ iterations; we upsample the images to 900×1200.

4.4. Comparison of Three Attention Nets

We compare our detectors to the baseline FasterRCNN detector on the CityPersons validation set in table 2, and we can make the following observations:

Attention helps overall. While looking at the overall performance measure of MR on the **R+HO** set, all three methods with attention mechanism show

Detector	Attention guidance G	R		HO		R+HO	
		MR	Δ MR	MR	Δ MR	MR	Δ MR
FasterRCNN	-	**15.52**	-	64.83	-	41.45	-
FasterRCNN+ATT-self	self attention	20.93	-5.41 pp	58.33	+6.50 pp	40.83	+0.62 pp
FasterRCNN+ATT-vbb	vbb supervision	16.40	-1.12 pp	57.31	+7.52 pp	39.49	+1.96 pp
FasterRCNN+ATT-part	part detections	15.96	-0.44 pp	**56.66**	+8.17 pp	**38.23**	+3.23 pp
FasterRCNN+part	-	16.90	-1.38 pp	59.03	+5.80 pp	40.64	+0.81 pp

Table 2. Results of detectors using different attention networks on the CityPersons validation dataset. The baseline detector is FasterRCNN; Δ MR indicates the performance gain on each subset; bold highlights the best results on each subset.

Detector	Occl.	R	HO	R+HO
RPN+BF[29]	×	9.58	74.36	24.01
DeepParts [26]	✓	11.89	60.42	22.79
MS-CNN[4]	×	9.95	59.94	21.53
JL-TopS [33]	✓	10.04	49.18	19.22
FasterRCNN	×	**9.18**	57.58	20.03
FasterRCNN+ATT-vbb	✓	10.33	**45.18**	**18.21**

Table 3. Comparison to state-of-the-art detectors on the Caltech test set. Numbers indicate MR; the second column indicates whether the method is designed for occlusion handling; bold highlights the best results in each subset.

Detector	R+HO	Δ MR
FasterRCNN	35.64	-
FasterRCNN+ATT-part	33.84	+1.80 pp
SpatialPooling[24]	37.37	-
TA-CNN[27]	34.95	-
RPN+BF[29]	30.23	-

Table 4. Comparison to state-of-the art detectors on the ETH dataset. Numbers indicate MR.

some improvement to the FasterRCNN baseline, ranging from 1pp to 3pp. We also compare to the FasterRCNN+part detector, which directly uses the part detection heatmaps as additional features for classification. The gap between FasterRCNN+ATT-part and FasterRCNN+part demonstrates that our attention net is a more effective way of exploiting occlusion patterns from part detections.

Attention helps more for heavy occlusion cases. The gap given by attention networks becomes larger for the heavy occlusion cases, which are more challenging to detect. Especially, we notice the FasterRCNN+ATT-part detector achieves more than 8pp improvement.

External attention > self attention. With self attention, FasterRCNN+ATT-self obtains a 0.62 pp gain on the **R+HO** set, which is smaller than the other two using external attention guidance. We also notice that FasterRCNN+ATT-self drops more than 5pp on the reasonable subset, which indicates the model concentrates too much on the hard cases, resulting in a limited ability to handle different occlusion levels. In contrast, FasterRCNN+ATT-vbb and FasterRCNN+ATT-part improve overall while obtaining comparable performance on the reasonable subset.

4.5. Generalization to Other Datasets

In order to investigate the generalization ability of the proposed methods, we also implement experiments on another two datasets: Caltech and ETH.

The results on the Caltech test set are shown in table 3, where we make a comparison to state-of-the-art methods. First, we can see MS-CNN [4], RPN+BF [29] and FasterRCNN achieve top results on the reasonable subset, but fail miserably on heavy occlusion cases due to the lack of occlusion handling. Our detector outperforms the previous state-of-the-art detector JL-TopS [33] by 4pp on the heavy occlusion subset, and establishes a new state of the art on the **R+HO** subset, which consists of a wide range of occlusion levels. We also show some qualitative results in figure 7, where we can see our detector produces robust detections for different occlusion patterns. For instance, in the first example containing crowds, people are occluded with each other, the other two detectors either miss some of them or produce many false positives, while our detector generates well-aligned detections for all of them.

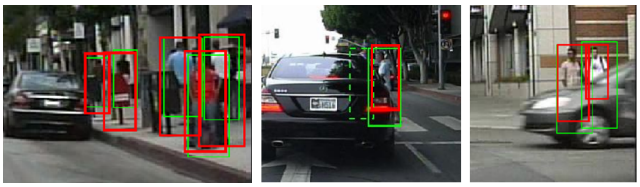
We apply our CityPersons models on the ETH dataset. Since no visible boxes are available, we can only evaluate on all occlusion levels. We show our results in table 4, where we can see that our attention model outperforms the FasterRCNN baseline by 1.80 pp. Compared to other state-of-the-art methods, the only one surpassing ours is RPN+BF [29]. In principle, our attention net can be added on top of any CNN based method. In this paper, we show the improvement to FasterRCNN, and we also expect similar behaviour if applied to RPN+BF.

The above results demonstrate that our attention models are robust to occlusion across different datasets, which are recorded in different cities, weather and illumination conditions, and involve various occlusion patterns.

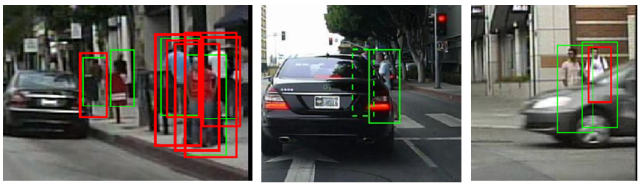
4.6. Discussion

In order to understand how attention handles occlusion in our models we analyze how Ω varies for pedestrian proposals with different occlusion patterns and different channels.

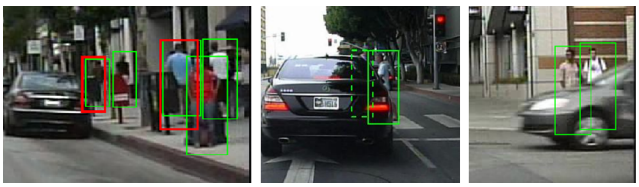
Assume we have H channels at the top convolutional layer, then Ω for proposal l is a vector of length H:



(a) FasterRCNN+ATT-vbb



(b) JL-TopS [33]



(c) MS-CNN [4]

Figure 7. Qualitative results from our detector and other competitive methods on the Caltech test set (at FPPI=0.1). The green solid and green dotted boxes indicate ground truth and ignored ground truth annotations; the red boxes denote detection results.

Figure 8. Visualization of how Ω behaves for occluded people. Each row shows the channel features on the left, and proposals (for six people detected) with the decreasing ranking of Ω for this channel on the right. The two channels represent feet and upper body respectively. For those people whose feet are occluded, Ω for the feet channel is lower ranked; while Ω for the upper body channel is highly ranked.

$$\Omega_l = [\omega_l^0, \omega_l^1, ..., \omega_l^{H-1}], \tag{6}$$

where ω_l^t will be applied on t th channel for re-weighting operation. In our detectors, $H = 512$.

The elements in Ω_l are then sorted in an increasing order, so as to get the ranking vector:

$$R_l = [r_l^0, r_l^1, ..., r_l^{H-1}], \tag{7}$$

where r_l^0 indicates the index of channel with the lowest impact in the final feature pool, and vice versa.

We denote the rank of channel t for proposal l as C_l^t, and it can be defined as:

$$C_l^t = m, \quad \text{if} \quad r_l^m == t. \tag{8}$$

For channel t, if $C_l^t > C_v^t$, i.e. ω_l^t ranks higher than ω_v^t, then this channel plays a more important role for proposal l than proposal v.

In figure 8, we show two channels, representing the feet and the upper body respectively. And for each channel, we show the proposals for six people detected in the image, with decreasing C value side by side. Among all channels, the given channel has a higher impact on the proposals on the left than on the right. We can see that for those people whose feet are occluded, the feet channel has a relatively lower impact than those fully visible people; on the other hand, the upper body is visible for all six proposals, but it ranks higher for occluded ones, this is because other channels for invisible parts are ranked lower. In this way, Ω re-weights the channels and allows occluded people to generate a high confidence in the final feature pool by upweighing visible channels.

5. Conclusion

In this paper, we propose to employ channel-wise attention to handle occlusion for pedestrian detection. From the visualization, we find that many channel features are localizable and often correspond to different body parts. Motivated by these findings, we design an attention net to generate attention vectors for re-weighting the top convolutional channels. This attention net can be added as an additional component to any CNN based detector. We explore different attention guidances, and find that all improve performance for occluded cases while the most effective one is the one based on part detections.

We report experimental results on the CityPersons, Caltech and ETH datasets, and show significant improvements over the baseline FasterRCNN detector. In particular, on CityPersons, we achieve a significant improvement of 8pp on the heavy occlusion subset and on Caltech, we outperform the previous state of the art by 4pp for heavily occluded people. Encouraged by the above results, we believe that the proposed method will also improve results for the general object detection task, where occlusion is also a major challenge.

Acknowledgments

This work was partially supported by the National Science Fund of China under Grant Nos.61702262, U1713208 and 61472187, the 973 Program No.2014CB349303, Program for Changjiang Scholars, and "the Fundamental Research Funds for the Central Universities" No.30918011322.

References

[1] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017. 2

[2] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, 2016. 2

[3] R. Benenson, M. Omran, J. Hosang, , and B. Schiele. Ten years of pedestrian detection, what have we learned? In *ECCV, CVRSUAD workshop*, 2014. 1

[4] Z. Cai, Q. Fan, R. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*, 2016. 2, 7, 8

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 6

[6] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34(4):743–761, 2012. 6

[7] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *CVPR*, 2010. 1, 2

[8] A. Ess, B. Leibe, K. Schindler, and L. V. Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, 2008. 6

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2

[10] A. Gonzalez-Garcia, D. Modolo, and V. Ferrari. Do semantic parts emerge in convolutional neural networks? *IJCV*, 126(5):476–494, 2017. 2

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[12] J. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians. In *CVPR*, 2015. 1, 2

[13] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv*, 2017. 2, 5

[14] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016. 3, 5, 6

[15] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015. 2

[16] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2

[18] J. Li, X. Liang, S. Shen, T. Xu, and S. Yan. Scale-aware fast r-cnn for pedestrian detection. *arXiv*, 2016. 2

[19] M. Mathias, R. Benenson, R. Timofte, and L. Van Gool. Handling occlusions with franken-classifiers. In *ICCV*, 2013. 1, 2

[20] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2

[21] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, 2012. 1, 2

[22] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, 2013. 1, 2, 4

[23] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2

[24] A. v. d. H. S. Paisitkriangkrai, C. Shen. Strengthening the effectiveness of pedestrian detection. In *ECCV*, 2014. 7

[25] M. Simon, E. Rodner, and J. Denzler. Part detector discovery in deep convolutional neural networks. In *ACCV*, 2014. 2

[26] Y. Tian, P. Luo, X. Wang, and X. Tang. Deep learning strong parts for pedestrian detection. In *ICCV*, 2015. 1, 2, 4, 7

[27] Y. Tian, P. Luo, X. Wang, and X. Tang. Pedestrian detection aided by deep learning semantic tasks. In *CVPR*, 2015. 7

[28] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 2

[29] L. Zhang, L. Lin, X. Liang, and K. He. Is faster rcnn doing well with pedestrian detection. In *ECCV*, 2016. 7

[30] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? In *CVPR*, 2016. 1, 2

[31] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. Towards reaching human performance in pedestrian detection. *PAMI*, 40(4):973–986, 2018. 1

[32] S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*, 2017. 1, 2, 3, 6

[33] C. Zhou and J. Yuan. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *ICCV*, 2017. 1, 2, 4, 7, 8