Taking a Deeper Look at Pedestrians
Supplementary material

Jan Hosang Mohamed Omran Rodrigo Benenson Bernt Schiele

Max Planck Institute for Informatics
Saarbrücken, Germany
firstname.lastname@mpi-inf.mpg.de

1. CifarNet training, the devil is in the details

Training neural networks is sensitive to a large number of parameters, including the learning rate, how the network weights are initialised, the type of regularisation applied to the weights, and the training batch size. It is difficult to isolate the effects of the individual parameters, and the best parameters will largely depend on the specific setup. We report here the parameters we used.

We train CifarNet via stochastic gradient descent (SGD) with a learning rate of 0.005, a momentum of 0.9, and a batch size of 128. After 60 epochs, we reduce the learning rate by a factor of 0.1 and train for an additional 10 epochs. Reducing the learning rate even further did not improve the classification accuracy. The other learning rate policies we explored yielded inferior performance (e.g. gradually reducing the learning rate each training iteration). Careful tuning of the learning rate while adjusting the batch size was critical.

Other than the softmax classification loss, the training loss includes a L2 regularisation of the network weights. In the objective function, this regularization term has a weight of 0.005 for all layers but the last one (softmax weights), which receives weight 1. This parameter is referred in `Caffe` as "weight decay".

The network weights are initialised by drawing values from a Gaussian distribution with standard deviation $\sigma = 0.01$, with the exception of the first layer, for which we set $\sigma = 0.0001$.

2. Grid search around CifarNet

Table 1 shows the detection quality of different variants of CifarNet obtained by changing the number and size of the convolutional filters of each layer. See related section 4.3.1 of the main paper. Since different training rounds have different random initial weights, we train four networks for each parameter set and average the results. We report both mean and standard deviation of the miss rate on our validation set.

We observe that using either too small or too large filter sizes throughout the network hurts quality. The network width also seems to matter, a network too narrow or too wide can negatively impact classification accuracy. All and all the "middle-section" of the table shows only small fluctuations in miss-rate (specially when considering the variance).

In addition to filter size and layer width, we also experimented with different types of pooling layers (max-pooling versus mean-pooling), see figure 2 of main paper. Other than on the first layer, replacing mean-pooling with max-pooling hurts performance.

The results of table 2 indicate that there is no set of parameters close to CifarNet with a clear advantage over the default CifarNet parameters. When going too far from CifarNet parameters, classification accuracy plunges.

3. Grid search for AlexNet

Table 2 presents the swipe of parameters used to construct the "Best parameters" entries in table 8 of the main paper. We vary the criterion to select negative samples and the SVM regularization parameters. Defaults are parameters are IoU < 0.5, and C $= 10^{-3}$.

Overall we notice that neither parameter is very sensitive ($1 \sim 2$ percent points fluctuations). When C is far from optimal large degradation is observed (10 per cent points). As seen in table 8 of the main paper the gap between default and tuned parameters is rather small ($1 \sim 2$ percent points).

4. Datasets statistics

In figure 1 we plot the height distribution for pedestrians in Caltech and KITTI training sets. Although the datasets are visually similar, the height distributions are somewhat dissimilar (for reference ImageNet and Pascal distributions are more look alike among each other).

It was shown in [1] that models trained in each dataset, do not transfer well across each other (compared to models trained on the smaller INRIA dataset).

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

Sizes \ # filters	$16, 16, 16$	$\mathit{32}, \mathit{32}, \mathit{64}$	$32, 64, 32$	$64, 32, 32$	$32, 32, 32$	$64, 64, 64$	$64, 32, 16$	Mean
$3, 3, 3$	48.4 ± 1.7	44.4 ± 1.0	43.6 ± 0.8	45.1 ± 1.1	45.2 ± 0.7	42.3 ± 1.3	46.6 ± 2.1	45.1
$\mathit{5}, \mathit{5}, \mathit{5}$	42.7 ± 4.2	41.1 ± 1.3	39.1 ± 1.0	38.9 ± 1.5	37.8 ± 1.6	38.3 ± 2.5	38.5 ± 1.3	39.5
$7, 5, 3$	43.3 ± 2.9	38.7 ± 2.4	38.6 ± 2.1	38.8 ± 0.9	40.2 ± 2.0	37.9 ± 1.7	39.7 ± 0.7	39.6
$7, 5, 5$	43.5 ± 2.5	40.2 ± 0.9	40.8 ± 2.6	38.4 ± 0.9	40.8 ± 1.5	40.0 ± 0.4	41.7 ± 2.5	40.8
$7, 7, 5$	43.5 ± 2.7	41.6 ± 3.0	43.3 ± 6.1	40.5 ± 2.9	39.8 ± 2.5	47.3 ± 2.5	41.6 ± 2.0	42.5
Mean	44.3	41.2	41.1	40.4	40.8	41.2	41.6	

Table 1: Detection quality (MR$\%$) as the number of filters per layer (columns) and filter sizes per layer (rows). CifarNet parameters are highlighted in italic. (MR: log-average miss-rate on Caltech validation set).

neg overlap \ C	10^{-6}	$10^{-5.5}$	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}
0.3	36.01%	33.62%	32.30%	32.22%	32.04%	32.42%	32.24%	32.26%	32.40%
0.4	36.01%	33.72%	32.43%	32.09%	32.16%	32.33%	32.23%	32.30%	32.20%
0.5	36.07%	33.90%	32.51%	32.03%	32.18%	32.53%	32.20%	32.28%	33.15%
0.6	36.50%	33.96%	32.43%	32.19%	32.24%	32.45%	32.29%	33.06%	34.61%
0.7	36.55%	34.32%	32.36%	32.05%	32.15%	32.55%	32.82%	33.83%	36.13%

(a) layer fc7

neg overlap \ C	10^{-6}	$10^{-5.5}$	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}
0.3	37.16%	32.49%	32.01%	31.88%	32.03%	32.18%	32.50%	32.40%	32.48%
0.4	37.16%	32.54%	32.07%	31.89%	32.14%	31.92%	32.46%	32.51%	32.56%
0.5	37.41%	32.61%	32.17%	32.07%	32.04%	31.84%	32.57%	33.12%	33.18%
0.6	37.54%	32.68%	32.14%	32.12%	32.22%	31.90%	32.93%	34.02%	35.85%
0.7	38.06%	32.67%	32.10%	31.89%	32.23%	32.32%	33.92%	35.92%	38.72%

(b) layer fc6

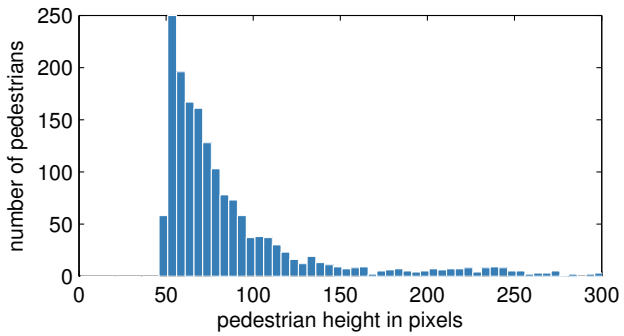
neg overlap \ C	10^{-6}	$10^{-5.5}$	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}
0.3	55.37%	36.77%	33.16%	32.75%	32.77%	33.29%	33.37%	34.28%	35.16%
0.4	55.89%	36.82%	33.17%	32.52%	32.82%	33.16%	32.79%	34.12%	35.42%
0.5	56.24%	37.09%	33.21%	32.65%	32.69%	33.14%	33.26%	34.95%	36.39%
0.6	56.68%	37.19%	33.40%	32.66%	32.83%	33.44%	34.17%	35.66%	38.28%
0.7	57.93%	37.60%	33.81%	32.85%	33.27%	34.23%	35.76%	38.98%	42.68%

(c) layer pool5

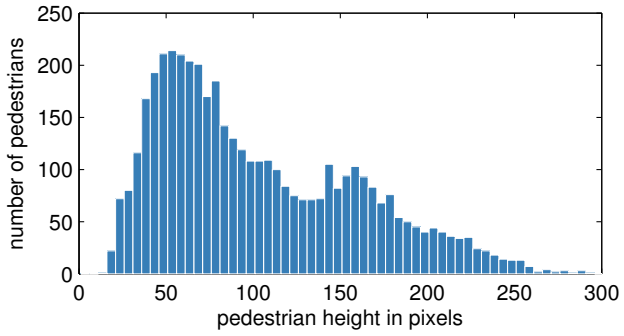
neg overlap \ C	10^{-6}	$10^{-5.5}$	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}
0.3	82.29%	64.90%	48.26%	44.67%	44.83%	43.66%	42.71%	43.36%	45.48%
0.4	82.29%	65.06%	48.66%	44.69%	44.67%	43.06%	42.41%	42.74%	44.81%
0.5	82.22%	65.23%	48.87%	44.68%	44.34%	42.98%	42.57%	43.30%	44.98%
0.6	82.22%	65.30%	48.69%	44.89%	44.39%	43.63%	42.92%	44.27%	46.35%
0.7	82.39%	65.96%	50.47%	45.62%	45.32%	44.86%	44.84%	46.31%	50.13%

(d) layer conv4

Table 2: Detection quality (MR) as function of the maximal IoU threshold to consider a proposal as negative example and the SVM regularization parameter C. (MR: log-average miss-rate on Caltech validation set)



(a) Caltech Reasonable training set



(b) KITTI training set

Figure 1: Histogram of pedestrian heights in different datasets.

5. Proposals statistics

In figures 2 and 3 we show statistics of different detectors on the Caltech test set, including the ones we use as proposals in our experiments. These figures complement table 9 of the main paper.

Our initial experiments indicated that it is important to keep a low number of average proposals per image in order to reduce the false positives rate (post re-scoring). This is in contrast to common practice when using class-agnostic proposal methods, where using more windows is considered better because they provide higher recall [2]. We filter proposals via a threshold on the detection score.

As can be seen in figure 2 a recall higher than 90% can be achieved with only $\sim$3 proposals per image on average (for Intersection-over-Union threshold above 0.5, the evaluation criterion). The average number of proposals per image is quite low because most frames of the Caltech test set do not contain any pedestrian.

In figure 3 we show the number of false positives at different overlap levels with the ground truth annotations. The bump around 0.5 IoU, most visible for `SpatialPooling` and `LDCF`, is an artefact of the non-maximum suppression method used by each method. Both these method obtain high quality detection, thus they must assign (very) low-

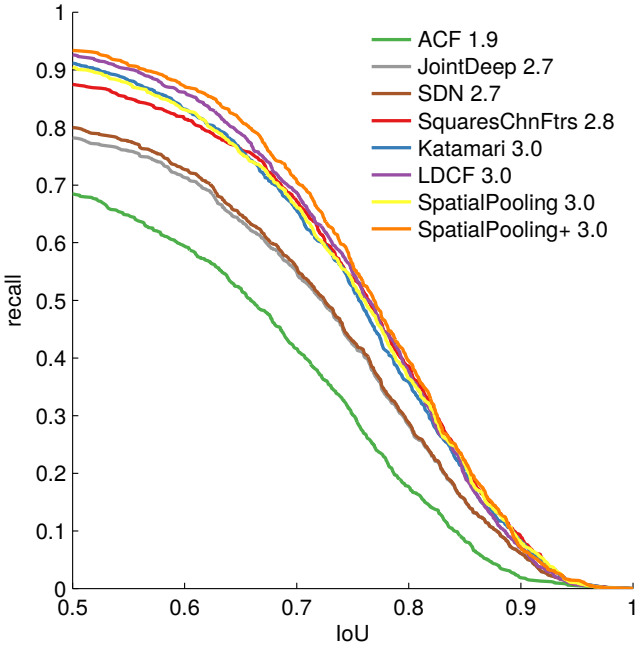


Figure 2: Recall of ground truth versus IoU threshold, for a selection of detection methods. The curves are cumulative distributions. The detections have been filtered by score to reach $\sim$3 proposals per image on average (number indicated in the legend).

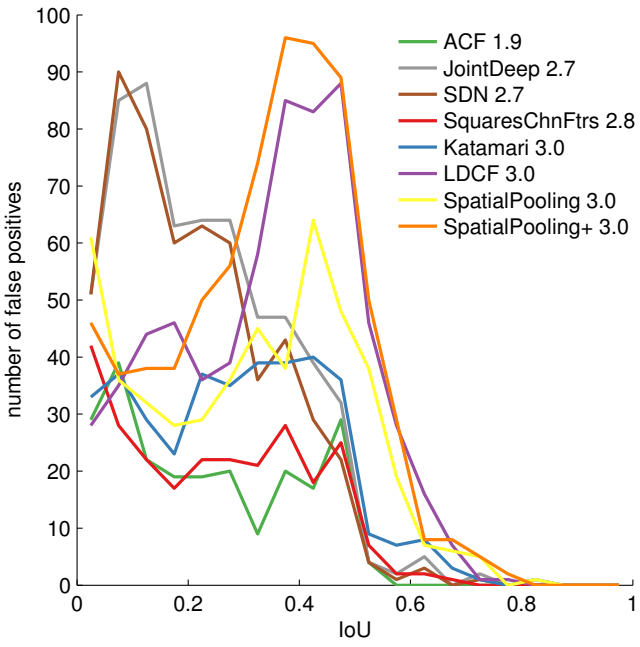


Figure 3: Distribution of overlap between false positives and ground truth of those false positives that do overlap with the ground truth. The curves are histogram with coarse IoU bins. Number in the legend indicates the average number of proposals per image (after filtering to reach $\sim$3).

scores to these false positives windows. To further improve quality the re-scoring method must do the same.

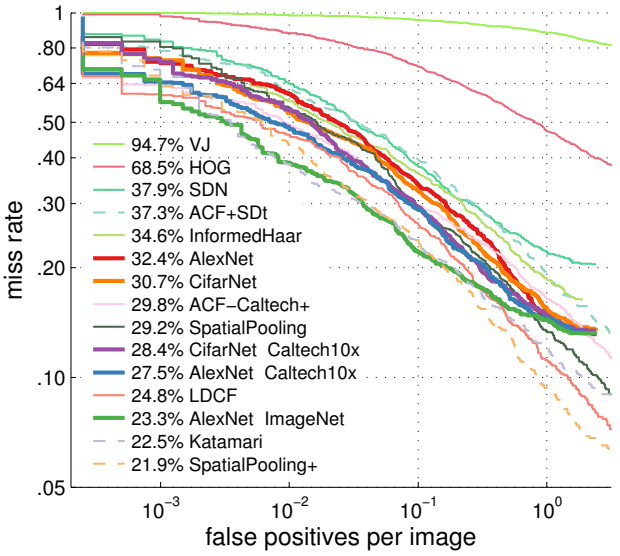
When using a method for proposals one desires to have high recall with high overlap with the ground truth (figure 2), as well has having false positives with low overlap with the ground truth (figure 3). False positive proposals overlapping true pedestrians will have pieces of persons, which might confuse the re-scoring classifier. Classifying fully centred persons versus random background is assumed to be easier task.

In table 9 of the main paper we see that AlexNet reaches top detection quality by improving over `LDCF`, `SquaresChnFtrs`, and `Katamari`.

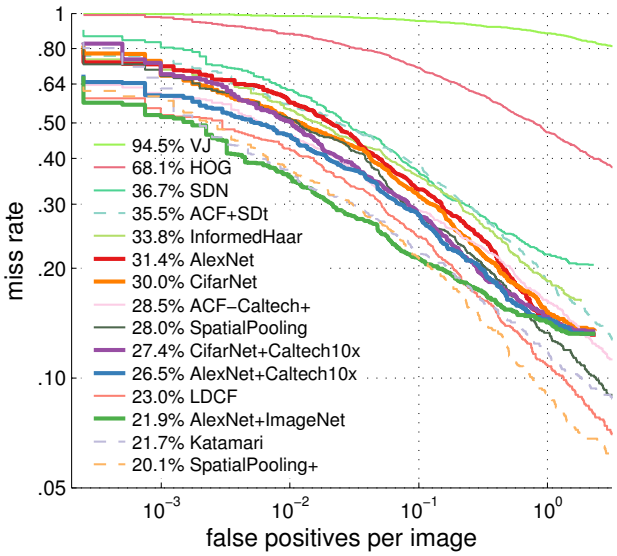
6. Error analysis

In the error analysis in the paper, we mention that the highest scoring false positives of our best models consist of localization errors of the detection proposals, of the AlexNet, but also of the ground truth. This suggests, that the detector struggles to rank partial pedestrians low enough: below low scoring, but well localized pedestrians. In this section and in figure 4 we present an experiment, which shows that fixing localization errors does not improve the performance by more than 2% log-average miss-rate, and thus most performance is lost by false detections on background.

Figure 4a shows the performance of our key results and other published results on the Caltech test set. For each of those pedestrian detectors we run the evaluation to determine which detections are false positives and remove those, which overlap with annotated pedestrians. For the remaining detections we plot the performance, as shown in figure 4b. This filtering step generously removes all detections of partial pedestrians that are counted as mistakes and only leaves false detections on background. All considered methods gain no more than 2% log-average miss-rate, which means that the conclusion from looking at the highest scoring false positives is wrong. Fixing localization issues will not have a great impact for any of the detectors. Instead *all methods struggle to distinguish pedestrians from background* and this problem has to be addressed to achieve bigger improvements under the current evaluation metric.



(a) Regular detection performance.



(b) Detection performance after removing all false positives that have any overlap with annotated pedestrians.

Figure 4: Performance of our key results (thick lines) and published methods on Caltech test set before and after removal of false positives that touch annotated pedestrians. Methods using optical flow are dashed.

References

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

[2] J. Hosang, R. Benenson, and B. Schiele. How good are detection proposals, really? In *BMVC*, 2014. 3