# How Far are We from Solving Pedestrian Detection? Supplementary material

Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang and Bernt Schiele
Max Planck Institute for Informatics
Saarbrücken, Germany
firstname.lastname@mpi-inf.mpg.de

## 1. Content

This supplementary material provides a more detailed view of some of the aspects presented in the main paper.

- Section 2 gives details of the `RotatedFilters` detector we used for our experiments (section 2.2 in main paper).
- Section 3 provides the detailed curves behind the summary bar plots for different test set subsets (see figure 3 and section 3.1 in main paper).
- Section 4 shows examples for each error type from the analysed detector, discusses the scale, blur and contrast evaluations, and revisits the oracle cases experiments in more detail (section 3.2 in main paper).
- Section 5 shows examples of how the new training annotations improve over the original ones (section 3.3 in main paper).
- Section 6 discuss the impact of new annotations on the evaluation of existing methods (MR ranking and recall-versus-IoU curves) (section 4.1 in main paper).
- Section 7 shows the effects of automatically aligning $10\times$ data with $1\times$data (section 4.1 in main paper).
- Figure 16 summarises our final detection results both in original and new annotations.

Other than this text we provide as supplementary material an annotations inspection tool described in section 5.1.

## 2. Rotated filters detector

For our experiments we re-implement the filtered channel feature `Checkerboards` detector [5] using the `LDCF` [4] codebase. The training procedure turns out to be slow due to the large number of filters (61 filters per channel). To accelerate the training and test procedures, we design a small set of 9 filters per channel that still provides good performance. We call our new filtered channel feature detector; `RotatedFilters` (see figure 1d).

The rotated filters are inspired by the filterbank of `LDCF` (obtained by applying PCA to each feature channel). The first three filters of `LDCF` of each features channel are the constant filter and two step functions in orthogonal directions, with the particularities that the oriented gradient channels also have rotated filters (see figure 1b). Our rotated filters are stylised versions of `LDCF`. The resulting `RotatedFilters` filterbank is somewhat intuitive, while filters from `Checkerboards`, are less systematic and less clear in their function (see figure 1c).

To integrate richer local information, we repeat each filter per channel over multiple scales, in the same spirit as `SquaresChnFtrs` [1] (figure 1a).

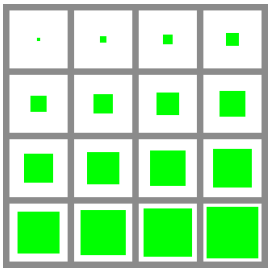
On the Caltech validation set, `RotatedFilters` obtains 31.6% $\text{MR}^{O}_{-2}$ using one scale (4x4); and 28.9% $\text{MR}^{O}_{-2}$ using three scales (4x4, 8x8 and 16x16). Therefore, we select this 3-scale structure in our experiments. On the test set, the performance of `RotatedFilters` is 19.2% $\text{MR}^{O}_{-2}$, i.e. a less than 1% loss with respect to `Checkerboards`, yet it is ~6x faster at feature computation.

In this paper, we use `RotatedFilters` for all experiments involving training a new model.

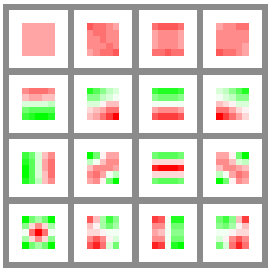
## 3. Results per test subset

Figure 2 contains the detailed curves behind figure 3 in the main paper ("subsets bar plot"). We can see that `Checkerboards` and `RotatedFilters` show good performance across all subsets. The few cases where they are not top ranked (e.g. figures 2e and 2h) all methods exhibit low detection quality, and thus all have similarly poor scores.

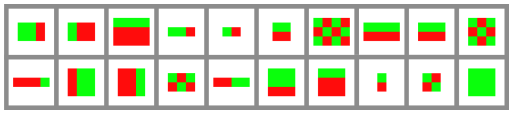
Figure 2 shows that `Checkerboards` is not optimised for the most common case on the Caltech dataset, but instead shows good performance across a variety of situations; and is thus an interesting method to analyse.



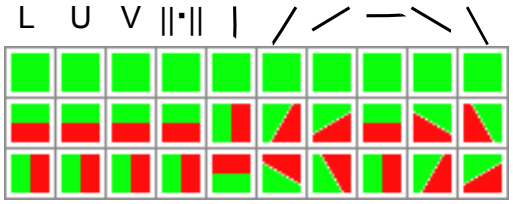
(a) `SquaresChntrs` [1] filters



(b) Some of the `LDCF` [4] filters. Each column shows filters for one channel.

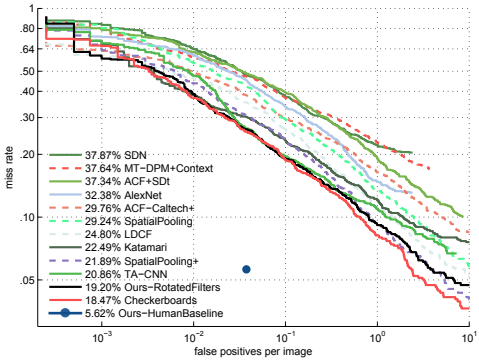


(c) Some examples of the 61 Checkerboards filters (from [5])

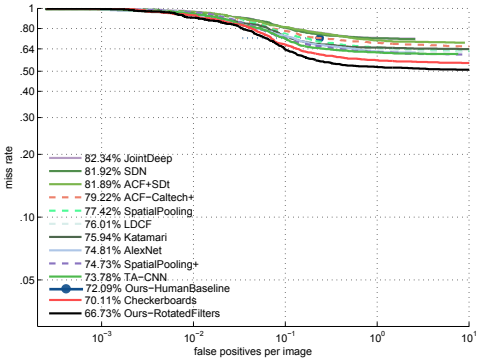



(d) Illustration of Rotated filters applied on each feature channel

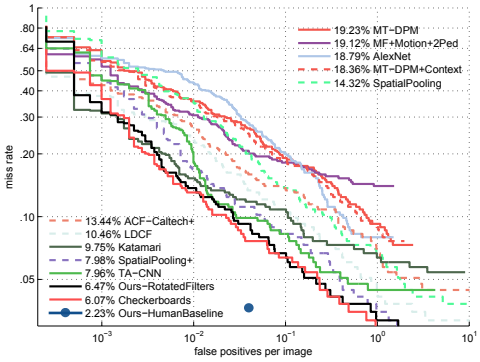
Figure 1: Comparison of filters between some filtered channels detector variants.



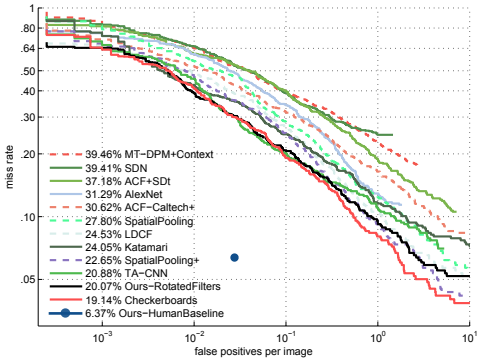

(a) Reasonable setting (IoU >= 0.5)



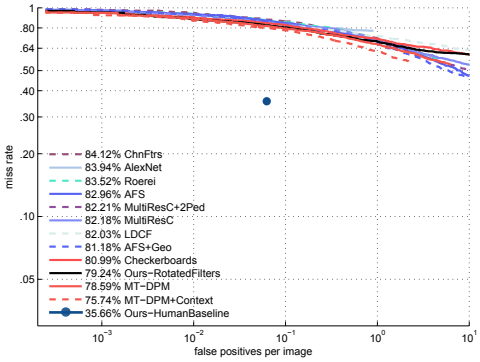

(b) Reasonable setting (IoU >= 0.8)



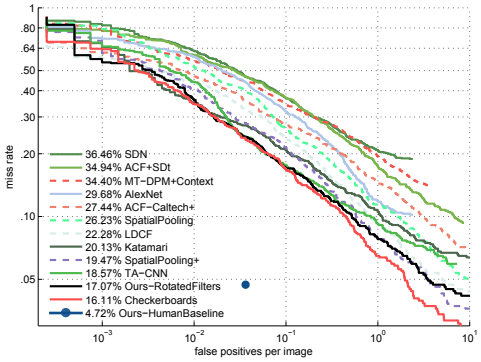

(c) Pedestrians larger than 80px in height



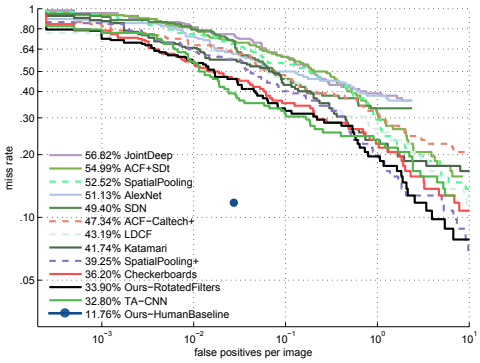

(d) Pedestrian height between 50px and 80px



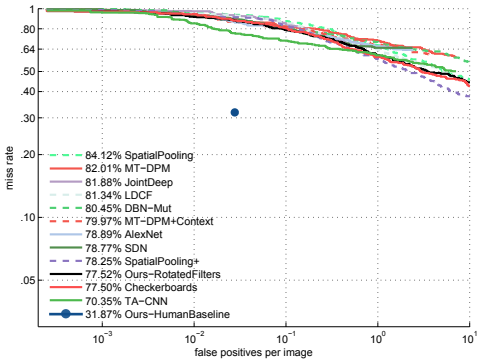

(e) Pedestrian height between 30px and 50px




(f) Non-occluded pedestrians




(g) Pedestrians occluded by up to 35%




(h) Pedestrians occluded by more than 35% and less than 80%.

Figure 2: Detection quality of top-performing methods on experimental settings depicted in "subsets bar plot" figure in the main paper.

## 4. `Checkerboards` errors analysis

**Error examples** Figure 7, 8, 9 and 10, show four examples for each error type considered in the analysis of the main paper (for both false positives and false negatives).

**Blur and contrast measures** To enable our analysis regarding blur and contrast, we define two automated measures. We measure blur using the method from [3], while contrast is measured via the difference between the top and bottom quantiles of the grey scale intensity of the pedestrian patch.
Figures 5 and 6 show pedestrians ranked by our blur and contrast measures. One can observe that our quantitative measures correlate well with the qualitative notions of blur and contrast.

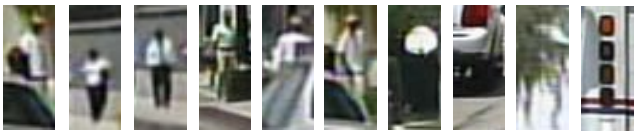
**Scale, blur, or contrast?** For false negatives, a major source of error is small scale, but we find small pedestrians are often of low contrast or blurred. In order to investigate the three factors separately, we observe the correlation between size/contrast/blur and score, as shown in figure 4. We can see that the overlap between false positive and true positive is equally distributed across different levels of contrast and blur; while for scale, the overlap is quite dense at small scale. To this end, we conclude that small scale is the main factor negatively impacting detection quality; and that blur and contrast are uninformative measures for the detection task.

### 4.1. Oracle cases

In figure 11, we show the standard evaluation and oracle evaluation curves for state-of-the-art methods. For the localisation oracle, false positives that overlap with the ground truth are not considered; for the background-versus-foreground oracle, false positives that do not overlap with the ground truth are not considered. Based on the curves, we have the following findings:

- All methods are significantly improved in each oracle evaluation.
- The ranking of all methods stays relatively stable in each oracle case.
- In terms of $\text{MR}^{\text{O}}_{-4}$, the improvement is comparable for localisation or background-versus-foreground oracle tests; the detection performance can be boosted by fixing either problem.

We also show some examples of objects with similar scores in figure 3. In both low-scoring and high-scoring groups, we can see both pedestrians and background objects, which shows that the detector fails to rank foreground and background adequately.



(a) Low-scoring objects

(b) High-scoring objects

Figure 3: Failure cases of Checkerboards detector [5]. Each group shows image patches of similar scores: some background objects are of high scores; while some persons are of low scores. We aim to understand when the detector fails through analysis.

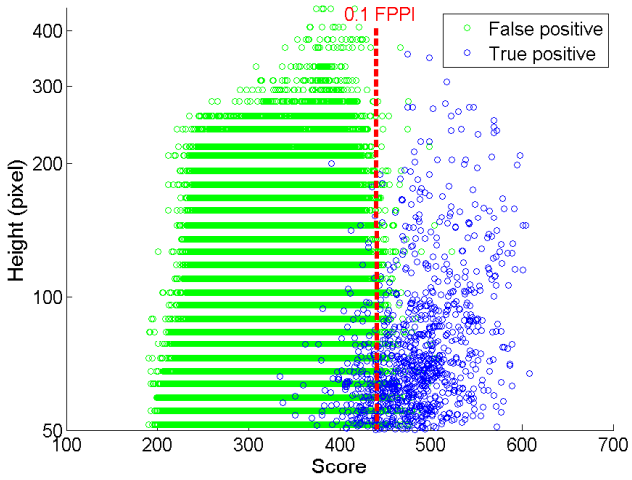
### 4.2. Log scale visual distortion

In the paper we show results for so called oracle experiments that emulate the case in which we do not make one type of error: we remove either mistakes that touch annotated pedestrians (localisation oracle) or mistakes that are located on background (background oracle).

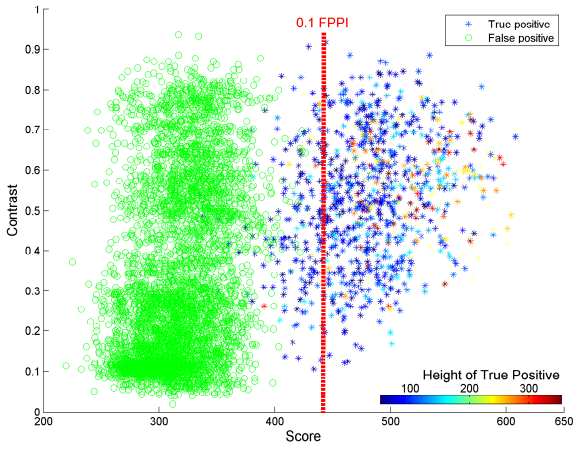
It is important to note that these are the only two types of false positives. If we remove both types the only mistakes that remain stem from missing recall and the result would be a horizontal line with very low miss rate.

Because of the double log scale in the performance plots on Caltech the curves look like both oracles improve performance slightly but the bulk of mistakes arise from a different type of mistakes, which is not the case.

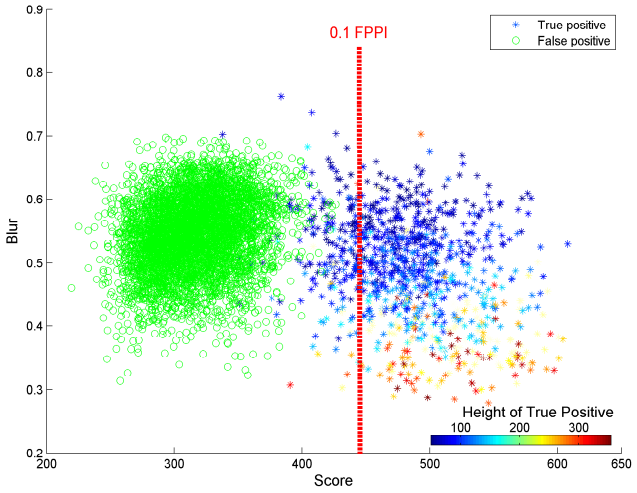
In figure 12 we illustrate how much double log scales distort areas. We often think of the average miss rate as the area under the curve, so we colour code the false positives in the plots by their type: the plot shows the ratio between localisation (blue) and background (green) mistakes at every point on the miss rate, but also for the entire curve. Both curves, 12b and 12c show the same data with the only difference that one shows localisations on the left and the other one on the right. Due to the double log scale, the error type that is plotted on the left seems to dominate the metric.




(a) Size versus score




(b) Contrast versus score




(c) Blur versus score

Figure 4: Correlation between size/contrast/blur and score.

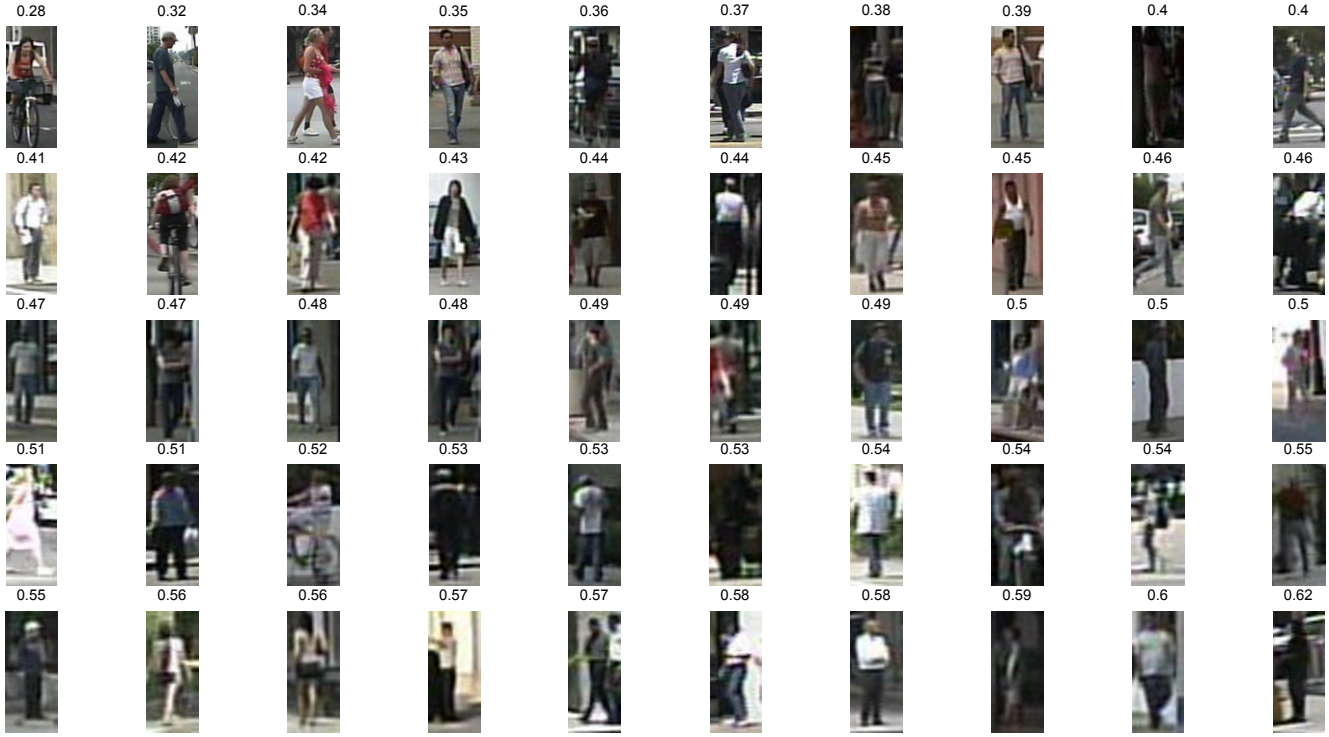



Figure 5: Examples for images with different levels of blur.

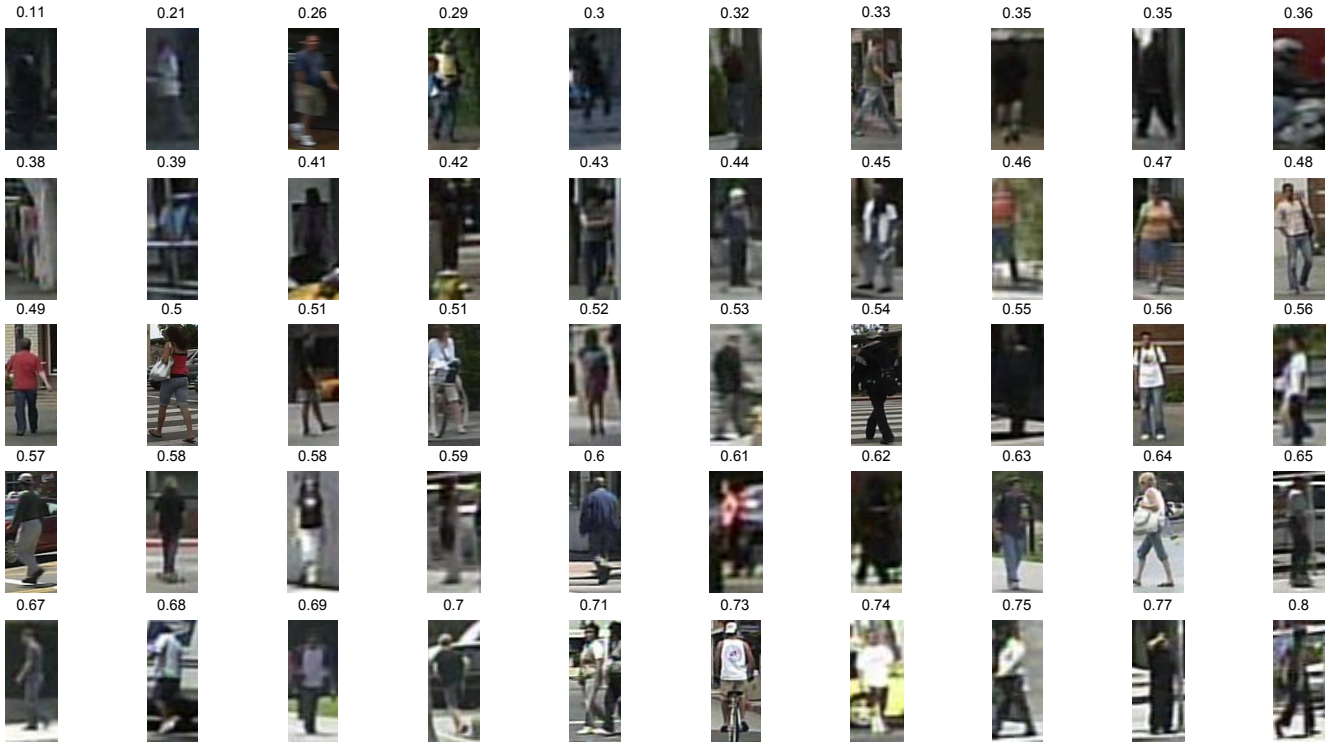
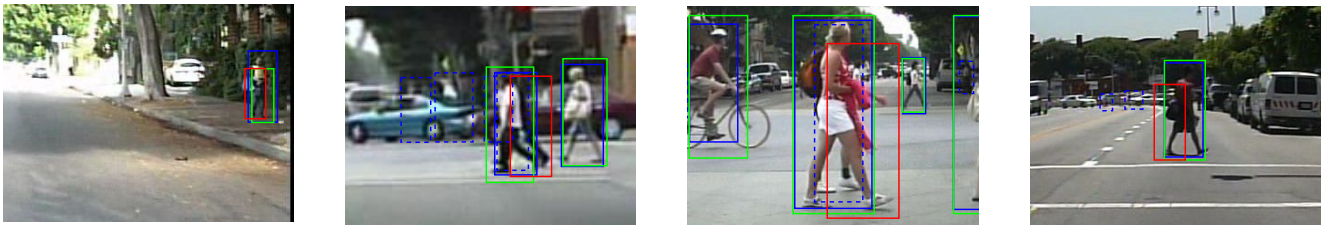



Figure 6: Examples for images with different levels of contrast.

(a) Double detection

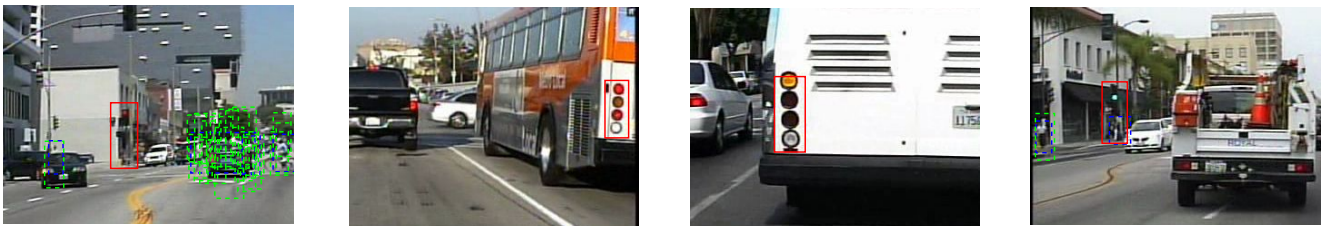


(b) Body parts

(c) Too large bounding boxes

Figure 7: Example localisation errors, a subset of false positives. False positives in red, original annotations in blue, ignore annotations in dashed blue, true positives in green, and ignored detections in dashed green (because they overlap with ignore annotations).

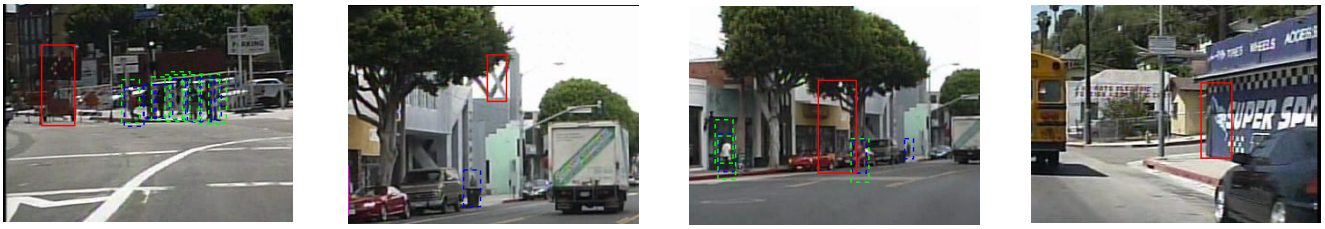
(a) Vertical structures



(b) Traffic lights

(c) Car parts

(d) Tree leaves



(e) Other background

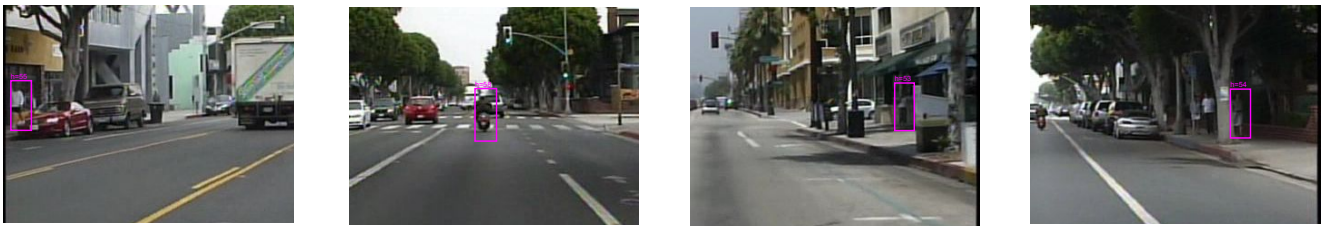
Figure 8: Example background errors, a subset of false positives. False positives in red, original annotations in blue, ignore annotations in dashed blue, true positives in green, and ignored detections in dashed green (because they overlap with ignore annotations).

(a) Fake humans

(b) Missing annotations

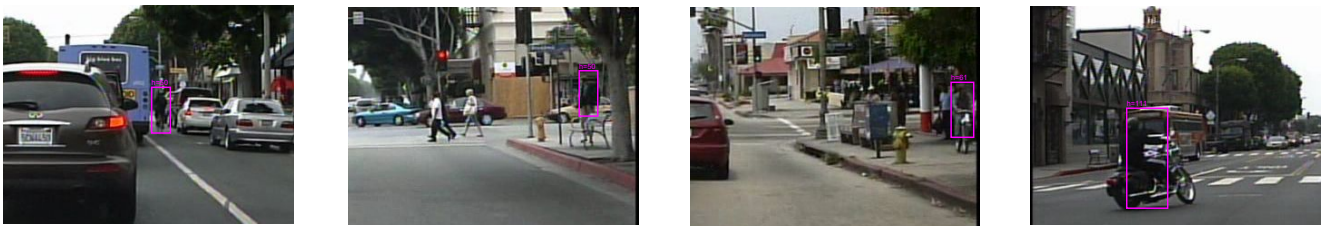
(c) Confusing

Figure 9: Example annotation errors, a subset of false positives. False positives in red, original annotations in blue, ignore annotations in dashed blue, true positives in green, and ignored detections in dashed green (because they overlap with ignore annotations).



(a) Small scale

(b) Side view



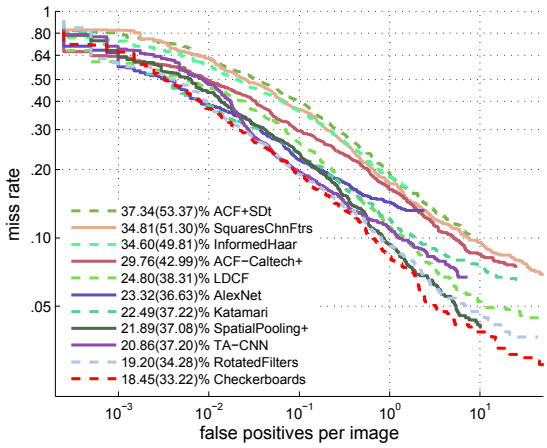
(c) Cyclists

(d) Occlusion

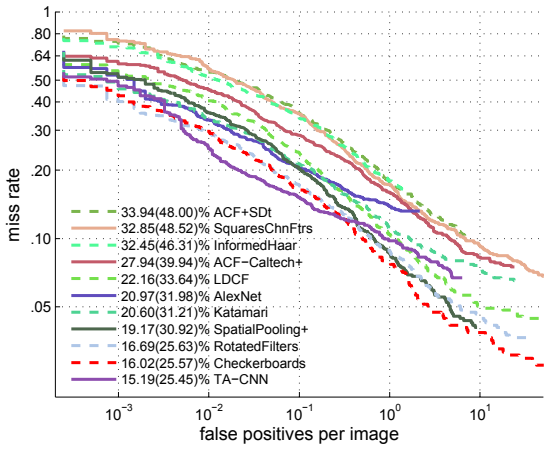
(e) Annotation errors

(f) Others

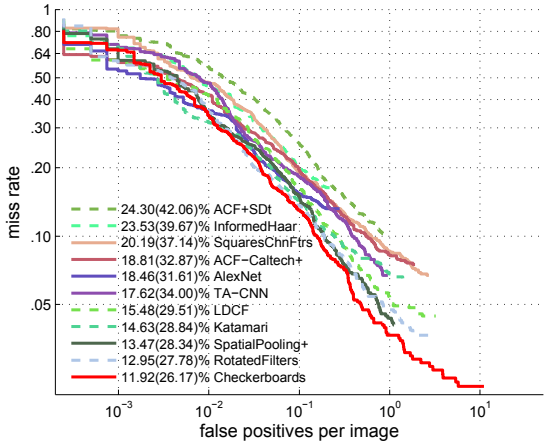
Figure 10: Example errors for different error types of false negatives. False positives in red, original annotations in blue, ignore annotations in dashed blue, true positives in green, and ignored detections in dashed green (because they overlap with ignore annotations).




(a) Standard evaluation (reasonable subset)

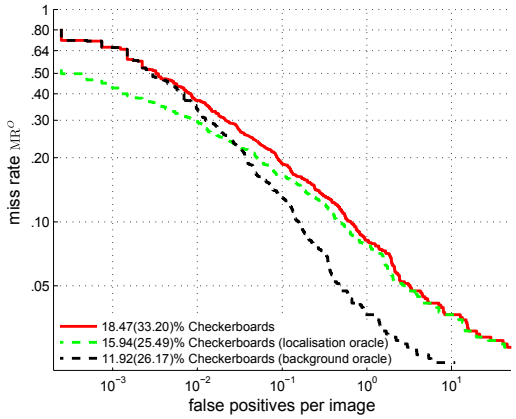



(b) Localisation oracle

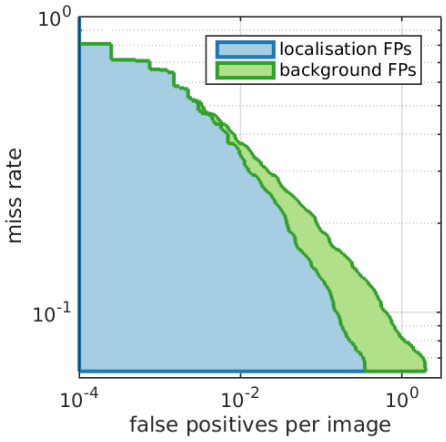



(c) Background-vs-foreground oracle

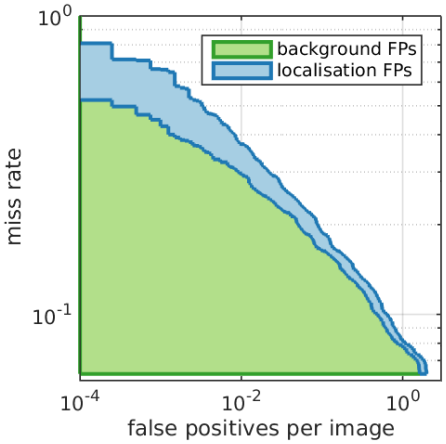
Figure 11: Caltech test set error with standard and oracle case evaluations. Both localisation and background-versus-foreground show important room for improvement. Both $\mathrm{MR}^{\mathrm{O}}_{-2}$ and $\mathrm{MR}^{\mathrm{O}}_{-4}$ are shown for each method at each evaluation.




(a) Original and two oracle curves for `Checkerboards` detector.




(b) Localisation FPs on the left.




(c) Background FPs on the left.

Figure 12: `Checkerboards` performance on standard Caltech annotations, when considering oracle cases . Localisation mistakes are blue, background mistakes green.

## 5. Improved annotations

In figure 13 we show original (red) and new annotations (green) on example frames from the test set. From the comparison, we can see that the new annotations are better aligned to the pedestrians. This results from the fact that head and feet are closer to the centre of the new bounding boxes.

### 5.1. Visualisation tool

We provide a Matlab visualisation tool (which includes part of Piotr Dollar's toolbox) to inspect the difference between original and new annotations. You can download the Caltech sequences from `http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/datasets/USA/` and run "`visualize_annotations('/path')`".

This tool will show the original and new annotations side by side. Instructions for how to run it can be found in "`GT-visualization/readme.txt`" in the supplementary material archive file.

Figure 13: Examples of differences between original (red) and new annotations (green). Ignore regions are drawn with dashed lines. These are the top 150 annotations, when sorted from smallest to largest IoU between original and new annotations.

## 6. Evaluation on original and new annotations

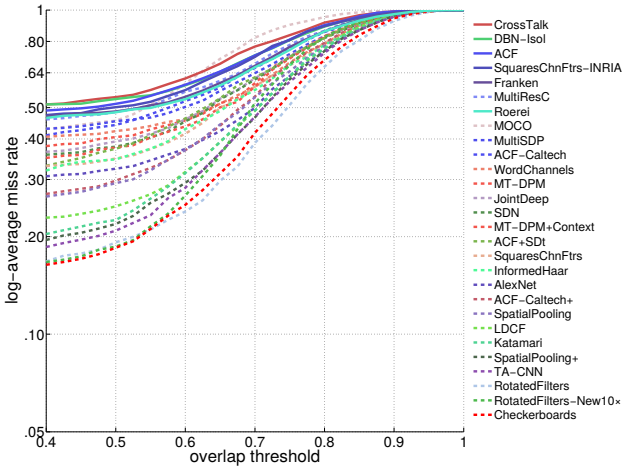
**Ranking** Figure 15 presents the ranking of all published Caltech methods up to CVPR 2015 when evaluated on $\text{MR}^{\text{O}}_{-2}$ (original annotations), or on $\text{MR}^{\text{N}}_{-2}$ (proposed new annotations). Although there are a few changes in ranking (e.g. `JointDeep` versus `SDN`), the overall trend is preserved. This is a good sign that the improved annotations are not a radical departure from previous ones. As discussed in the paper (and in other sections of the supplementary material), improved annotations matter most for future methods (going further down in MR), and for the low FPPI region of the curves (high confidence mistakes).

**`RotatedFilters`** Figures 16a and 16b show the results of our methods `RotatedFilters`, `RotatedFilters-New10x`, and `RotatedFilters-New10x-+VGG`; on the original and new annotations respectively. Using improved annotations during training (`-New10x`) does improve results both on original and new annotations.

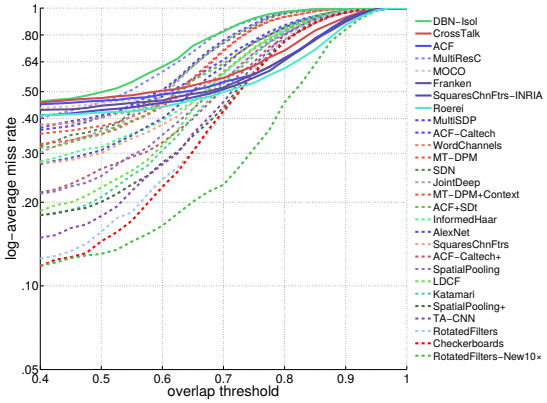
**MR versus IoU** Section 3.3 (and table 3) of the main paper discuss an empirical measure of how the new annotations are better aligned. Here we provide some more details. Figure 14 plots $\text{MR}^{\text{O}}_{-2}$ and $\text{MR}^{\text{N}}_{-2}$ of top performing methods versus the overlap criterion for accepting detections as true positives (IoU threshold). The standard evaluation uses IoU threshold 0.5. On these plots methods trained on INRIA have continuous lines, methods trained on Caltech dashed ones (see also figure 15).
In figure 14a (original annotations) the ranking of the methods remains stable as the overlap threshold becomes stricter (consistent with the observations in [2]). Interestingly we observe a different trend in figure 14b (new annotations).
When evaluating $\text{MR}^{\text{N}}_{-2}$(new annotations) we see that methods training on INRIA, albeit having a poor performance at IoU $= 0.5$, perform comparatively well at higher IoU, eventually overpassing all methods trained on raw Caltech data. We attribute this to the fact that INRIA training data is of better quality (better aligned training samples), and thus the detectors have learnt to localise better. This difference in trend between original and new annotations confirms that our improved annotations are better with respect to localization. Table 3 in the main paper provides a summarised version of figure 14.

## 7. Impact of aligning Caltech$10\times$

We can see from 14b that using our semi-automatically aligned Caltech $10\times$ training data provides a significant boost in localization quality. From `RotatedFilters` to `RotatedFilters-New10x` the $\text{MR}^{\text{N}}_{-2}$ improves across the full IoU range. Figure 17 shows qualitative results for the alignment procedure done over the $10\times$ training data.

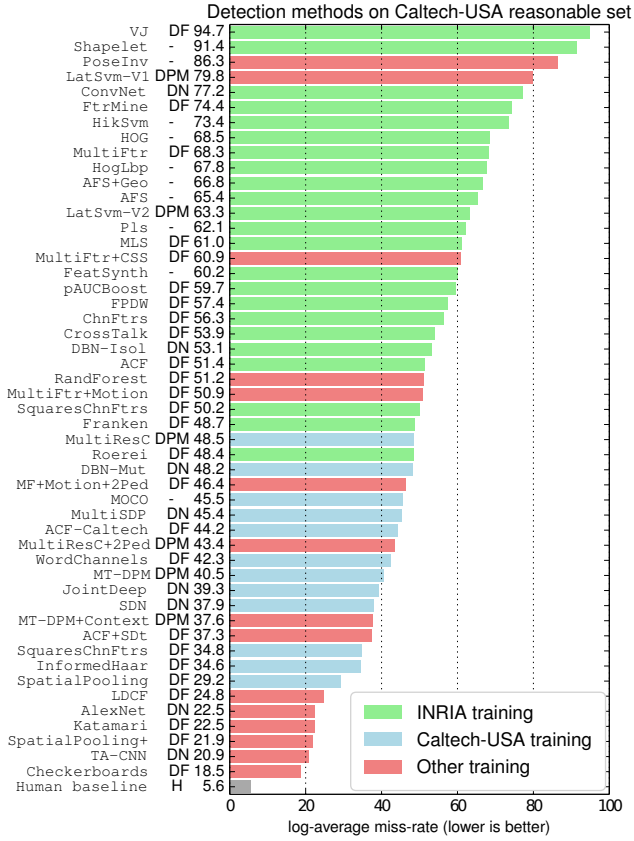



(a) Original annotations, $\text{MR}^{\text{O}}_{-2}$

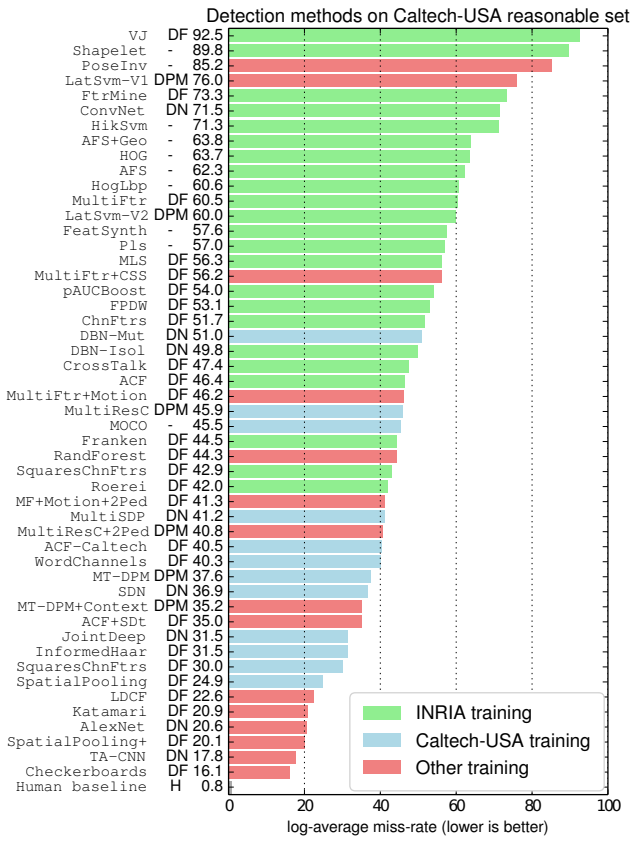



(b) New annotations, $\text{MR}^{\text{N}}_{-2}$

Figure 14: Plot of log-average miss rate versus overlap threshold (IoU) for the top-performing methods on the "reasonable" experimental setting. Methods trained on INRIA are represented with solid curves. On the new annotations, these behave better than methods trained on Caltech-USA original when we apply a stricter overlap criterion.

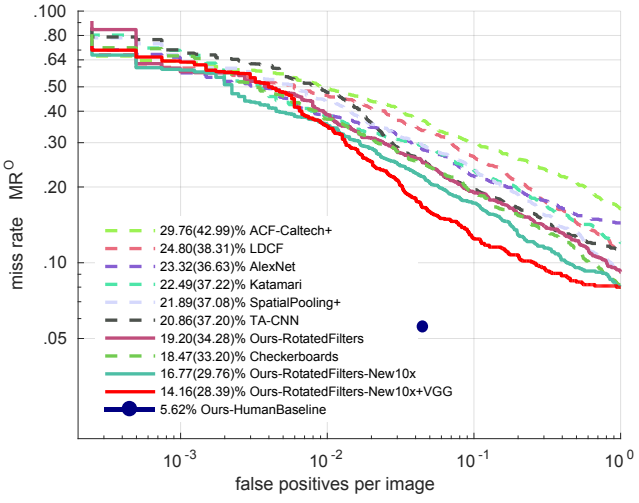



(a) $MR^{O}_{-2}$ Ranking of methods

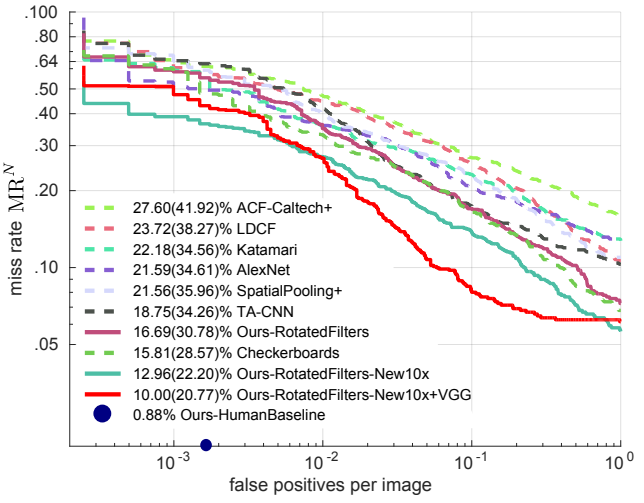



(b) $MR^{N}_{-2}$ Ranking of methods

Figure 15: Ranking of Caltech methods (CVPR 2015 snapshot) with original and new annotations. DF: decision forest, DPM: deformable parts model, DN: deep network.




(a) Evaluation on original annotations. Legend indicates $MR^{O}_{-2}(MR^{O}_{-4})$.




(b) Evaluation on new annotations. Legend indicates $MR^{N}_{-2}(MR^{N}_{-4})$.

Figure 16: Performance of top detectors evaluated on original and new annotations.

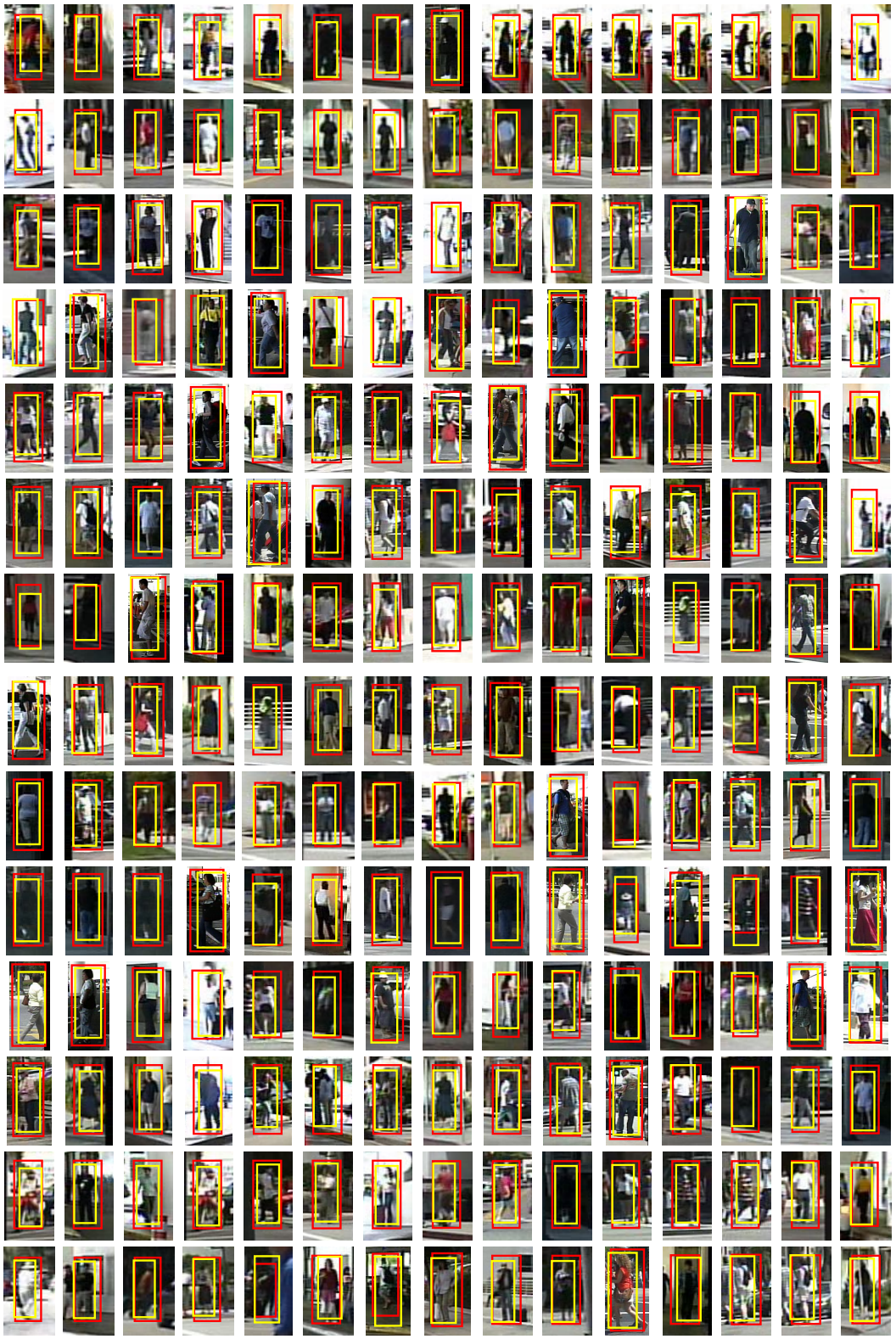


Figure 17: Examples of original annotations before (red bounding boxes) and after automatic alignment (yellow bounding boxes) using the `RotatedFilters` detector.

## 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. 2

[2] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 2012. 14

[3] Crete-Roffet F., Dolmiere T., Ladret P., and Nicolas M. The blur effect: Perception and estimation with a new no-reference perceptual blur metric. In *SPIE Electronic Imaging Symposium Conf Human Vision and Electronic Imaging*, May 2007. 4

[4] W. Nam, P. Dollár, and J. H. Han. Local decorrelation for improved detection. In *NIPS*, 2014. 2

[5] S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *CVPR*, 2015. 2, 4